Invited review: Assessing experimental designs for research conducted on commercial dairies¹

R. J. Tempelman²

Department of Animal Science, Michigan State University, East Lansing 48824-1225

ABSTRACT

Because of increasing constraints placed on conducting large studies at universities, more research is being conducted on commercial dairies, thereby raising some implications for experimental designs and data analysis. For example, experimental units are often specified to be pens of animals in on-farm studies, thereby requiring that at least 2 pens be used per treatment group in a single-dairy study. Even when treatments are compared within pens, the precision of inference on treatment differences is still primarily limited by the number of pens in the study, rather than the number of cows per treatment in each pen. Other challenges with on-farm studies include proper blocking and randomization of cows or pens to treatments. On the other hand, multiple farm studies are attractive, because they facilitate a broader scope of inference on treatment effects across a wider range of management or climatic conditions and genetic backgrounds compared with single-site university studies. Furthermore, studies based on multiple farms or multiple pens within a single large farm can facilitate greater power for treatment comparisons on binary reproduction or health responses than can be achieved at a smaller research herd. Because quantitative geneticists have been analyzing commercial dairy data for decades, they have developed useful data analysis techniques that should be harnessed to facilitate even greater statistical scope and power for on-farm studies.

Key words: experimental design, experimental unit, scope of inference

INTRODUCTION

Dairy cattle research conducted at university experiment stations has been and will continue to be vitally important in terms of providing knowledge that

can be readily used by the dairy industry. University herds are generally best suited for basic fundamental research geared toward determining various physiological mechanisms or modes of action that are important for dairy production (St-Pierre and Jones, 1999). This is particularly true when constant vigilance, intensive sampling, proximity to laboratories, and compliance to the experimental plan are vital.

Recently, various funding and environmental regulatory pressures have increasingly redirected more applied dairy research away from universities to commercial dairy venues. On-farm studies are often less expensive than applied research at universities on a per-herd basis; furthermore, they may be sometimes more relevant to modern production systems than studies conducted in smaller and often very differently managed research herds (St-Pierre and Jones, 1999). Nevertheless, there are several critical issues to consider when conducting on-farm research in commercial dairies. First, control of randomization of cows to treatments may be compromised as dependent upon the concerns of the farm manager, particularly when it is rather uncertain whether or not some treatments may have a detrimental effect on various aspects of dairy cattle performance. Second, and concomitant with increasing average herd sizes, a greater number of farms house and manage animals together in pens, thereby complicating issues for true experimental replication (St-Pierre, 2007).

Quantitative genetics researchers have been analyzing on-farm dairy data, primarily DHIA records, for many decades and have generally been quick to adopt cutting edge statistical tools for the analysis of such data. As the clearest example of this, the mixed model is the most commonly used statistical model for the analysis of many university and experimental research designs, with its development being particularly inspired by problems in dairy cattle breeding (Henderson, 1953; Henderson et al., 1959). In fact, it was within the context of these applications that optimal methods of statistical inference for the mixed model were proposed [i.e., an efficient algorithm for providing generalized least squares estimates of treatment effects and BLUP of random effects (Henderson et al., 1959; Goldberger, 1962)]. Mixed models do not only provide the backbone

Received May 30, 2008.

Accepted August 25, 2008.

¹Support from the Michigan Agricultural Experiment Station (Project MICL01822) is gratefully acknowledged.

²Corresponding author: tempelma@msu.edu

for virtually all genetic evaluation systems for livestock today, but generalized least squares is used in statistical software like SAS PROC MIXED/GLIMMIX (Littell et al., 2006) to properly analyze dairy research data.

It is the intent of this review to provide some direction for the design and analysis of on-farm dairy research, primarily providing an extensive power analysis for some common designs that involve pens or groups of animals, whether it be for continuous production responses or for binary health-reproduction responses. It will be particularly emphasized that different sizes of experimental units (cows versus pens versus herds) will determine the appropriate scope of inference for the on-farm study at hand. The potential utility of statistical model extensions, previously proposed by quantitative geneticists, for facilitating a broader scope of inference is discussed further as well.

EXPERIMENTAL DESIGN CONCEPTS

There are several basic principles that are essential for proper experimental design, whether they be applied to university research stations or commercial dairies. First, treatments should be randomly assigned to experimental units. When individual cows are randomly assigned to treatments, such that cows in the same treatment are not grouped or managed together, cows then serve as the experimental units as in these completely randomized designs (CRD). Furthermore, when the experiment can be conducted over several different time periods such that all cows receive all treatments but in a different time order or sequence, as in a Latin square or crossover design, then cows serve as a blocking factor along with time periods. It is well established that the statistical power of the crossover design generally exceeds that of the CRD for the same number of cows, as treatments are compared within cows in the crossover. It is important to note, however, that even before randomization for such studies, whether based on CRD or block designs, the cows chosen for the study are assumed to be a random or representative sample from a conceptually larger population of cows that represents the target population for inference (e.g., primiparous Holsteins).

When cows are grouped together in pens or paddocks as in large dairies, then the same randomization principles essentially apply except that now pens of cows, rather than individual cows themselves, represent the experimental units. In other words, it is essential that pens are randomly assigned to treatments in a CRD. Similarly, different time sequences of treatments are randomly assigned to each pen in a crossover study, just as they would for cows in individual cow studies. For example, if there are 2 treatments, A and B, to

compare within each pen across 2 time periods, then one-half of the pens would be randomly assigned to the A→B sequence (i.e., receiving A in the first period and B in the second period), whereas the other half of the pens would, by default, be assigned to the B→A sequence.

Of course, several experimental units per treatment are required to establish experimental error such that at least 2 pens must be randomly assigned to each treatment; however, we should be quick to note that 2 pens per treatment may not generally ensure sufficient statistical power, even if one could indefinitely increase the number of cows within each pen. Cox (1980) likely encountered many difficult statistical consulting sessions with dairy scientists when he writes: "One could merely assume that the effects of pens were negligible. This is equivalent to assuming that, when feeding a single diet, that the gains of two animals in the same pen are no more alike than the gains of two animals each in different pens." This temptation to then define individual cows as the experimental unit in pen studies has too often slipped into peer-reviewed literature, as also noted by St-Pierre and Jones (1999), such that Cox (1980) later bluntly retracts: "...pen-to-pen variation has been important too often to make such an assumption credible." The *Journal of Dairy Science* in its Instructions to Authors clearly stipulates that "The experimental unit is the smallest unit to which an individual treatment is imposed. For group-fed animals, the group of animals in the pen or the paddock is the experimental unit; therefore, groups must be replicated." It is important to further indicate that simply including pens as a source of variation in the model does not necessarily resolve this issue, particularly if pens are treated as fixed effects. If this approach is taken, then the scope of inference would be narrow (McLean et al., 1991); that is, any inferences derived from such an analysis would only apply to those specific pens used in the study. Further warning on biased estimates and type I error rates when treating individual cows as the experimental units for pen studies was also recently provided by St-Pierre (2007).

FACTORS INFLUENCING STATISTICAL POWER

A pressing issue for on-farm studies, in which pens are the experimental units, involves an appropriate choice for the numbers of pens and cows per pen. That is, how many pens or groups of animals per treatment should be considered when pens are the experimental units? Would it ever be worthwhile to conduct an experiment on one particular farm if there were only 2 or 3 pens per treatment?

To properly address this question, it is important to review what influences classical statistical power. Recall that statistical power is the probability that one would correctly conclude that 2 or more treatments have different mean responses. For the purposes of this review, we'll consider power to be sufficient if this probability exceeds 80%; however, depending on the context, anywhere between 70 to 99% power might be deemed acceptable or absolutely necessary.

Most researchers understand that the larger the true mean difference between treatment groups, the greater the statistical power will be. Power also depends on the specification of a desired type I error rate, denoted as α , which is the probability that one will conclude that 2 or more treatments have different mean responses when in fact there are no treatment effects. Typically, journals like the *Journal of Animal Science* or the *Journal of Dairy Science* dictate α to be less than or equal to 0.05, and certainly no greater than 0.10, for a researcher to be able to claim a statistically significant result. Beyond the true mean difference and the type I error rate, a third primary determinant of power is the number of experimental units. Again, depending upon the design and scope of the study, this may be defined by either the number of cows or pens or even the number of farms as described later.

However, statistical power vitally depends upon a fourth and perhaps least understood component, at least with respect to the design of pen studies. Variability from different sources of random effects as measured by the corresponding variance components (VC) may have the greatest effect on comparisons between various experimental designs for statistical power. As an aside, VC estimation has been a primary interest of dairy quantitative geneticists since Henderson (1953). First, there is the variability, σ_e^2 , between multiple measurements of the same response within cows, which is often characterized as the within-cow or measurement error VC; quantitative geneticists typically refer to σ_e^2 as the temporary environmental variance (Van Vleck, 1993). The VC characterizing variability between cows (σ_c^2) describes the biological or innate variability between cows and is generally referred to as the permanent environmental variance (Van Vleck, 1993). Let us further denote σ_p^2 as the VC measuring variability between pens and σ_h^2 as the VC measuring variability among herds (or among herds during a particular season within a particular year) as might be important in a multierd study. It is important to realize that the arrangement of the design factors, specifically cows, pens, herds, or all three, define the design structure for

the experiment. Elements of these design factors are typically specified to be random. That is, the cows, pens, or herds actually used for a particular study are believed to be a random subset of all cows, pens, or herds that could have been used for the study such that it is the general intent of the investigator that conclusions drawn from a particular experiment apply to all cows, pens, or herds from which these entities are considered to be representative of, not just to those that were actually used in the study.

There are other sources of variability due to interactions that are critical to specify in some on-farm experiments; in fact, by definition, treatment $\times$ design structure interactions define the nature of replication in experimental designs (Milliken and Johnson, 1984). For example, denote σ_{p*t}^2 as the VC due to pen $\times$ treatment interaction. That is, σ_{p*t}^2 can be used to express the VC of true treatment differences across pens such that if treatment differences are not truly variable across pens, then $\sigma_{p*t}^2 = 0$. Similarly, σ_{c*t}^2 can be used to define the VC due to cow $\times$ treatment interaction (i.e., the true treatment difference is variable or random across cows), whereas σ_{h*t}^2 defines the VC due to herd $\times$ treatment interaction (i.e., true treatment difference is variable across herds).

Depending on the design and scope of inference for the study as discussed later, any of the aforementioned VC could combine to determine the experimental variability and hence the statistical power for a particular study. Define the corresponding composite VC as σ_r^2 . Note that for all possible experimental designs, σ_r^2 is always a function of σ_e^2 ; furthermore, σ_r^2 may be a function of any of $\sigma_c^2, \sigma_{c*t}^2, \sigma_p^2, \sigma_{p*t}^2, \sigma_h^2$, and σ_{h*t}^2 for different sizes of experimental units (i.e., cows, pens, and perhaps even farms). Along these same lines, the number, n , of experimental units per treatment may be defined by cows, pens, or farms, again depending upon the design and scope of inference.

POWER DETERMINATIONS FOR CONTINUOUS RESPONSES

This section concentrates on power determinations for continuously recorded responses, like milk production or weight gain, and assumes that all such responses are normally distributed. Suppose that a study is conducted to compare 2 treatment groups, A and B, having respective unknown true means of μ_A and μ_B . Hence, the true unknown mean difference is $\delta = \mu_A -$

μ_B , and it is desired to have sufficient power to conclude that $H_1: \delta \neq 0$ if indeed $\delta \neq 0$. For a particular type I error rate of α , a specific number of experimental replicates (n) per treatment, and a particular design structure, the single most important function of δ and σ_r^2 influencing statistical power is the effect size:

$$\Delta = \frac{\delta}{\sigma_r}.$$

This term was first coined by Cohen (1988), where $\sigma_r = \sqrt{\sigma_r^2}$ defines the standard deviation between experimental units and may be a function of any number of constituent VC as previously noted. As a side note, for the joint comparison of $t > 2$ treatments having means $\mu_1, \mu_2, \dots, \mu_t$, the use of Δ might be replaced by a particular function of

$$\frac{\sqrt{\sum_{i=1}^t (\mu_i - \mu)^2}}{\sigma_r},$$

where $\mu = \frac{1}{t} \sum_{i=1}^t \mu_i$ (O'Brien and Lohr, 1984; Stroup, 2002). Once Δ is determined, it can be readily used to determine the noncentrality parameter that can be subsequently used to compute power. A particularly effective computational strategy for determining power for general experimental designs has been presented by Stroup (2002), meticulously demonstrated by Kononoff and Hanford (2006), and has been used by Tempelman (2004) for determining power for individual cow studies. This computational strategy is subsequently used for all power determinations presented in this paper as well.

On many university experiment research stations involving individually treated cows, Latin square designs are typically used with cow as 1 blocking factor and period as the second blocking factor. Hence, the number of cows defines the number of experimental replicates with $\sigma_r^2 = \sigma_e^2$ given that treatments are compared within cows. For the case of replicated 4×4 Latin squares and double crossover designs for individually treated cows, statistical power determinations were determined for various numbers of cows by Tempelman (2004), who noted that, despite the wide myriad of possible choices for δ or σ_e , statistical power is simply a function of their ratio $\Delta = \frac{\delta}{\sigma_e}$. The following sections

characterize statistical power for 3 common designs based on the use of pens of animals as being the experimental units.

CRD

Consider the power of test for a CRD where pen is the experimental unit; that is, a number (n) of pens are randomly assigned to each of 2 or more treatments within a herd. In this case,

$$\sigma_r^2 = \frac{\sigma_e^2 + \sigma_{c(t)}^2}{d} + \sigma_{p(t)}^2$$

for d being the number of cows per pen. Here, $\sigma_{c(t)}^2 = \sigma_c^2 + \sigma_{c* t}^2$ and $\sigma_{p(t)}^2 = \sigma_p^2 + \sigma_{p* t}^2$ refer to the VC characterizing the variability between cows within treatments and between pens within treatments, respectively. Note that here σ_r^2 defines the variability between the pen sample means; that is, the ANOVA expected mean square (**EMS**) for pens within treatments divided by the number of observations (d) on each pen or experimental unit. If the study is conducted within only 1 herd, then σ_r^2 does not depend upon σ_h^2 or $\sigma_{h* t}^2$ as indicated above; however, then the scope of inference is substantially limited to that herd as described later. For the situation where there is $d = 1$ cow per pen (i.e., individually fed cows as in tie-stall or research herd situations), then cow is essentially the experimental unit such that $\sigma_r^2 = \sigma_e^2 + \sigma_{c(t)}^2 + \sigma_{p(t)}^2$.

Figure 1a displays the power curves for the CRD design for different number of pens per each of 2 treatments and for different values of Δ ranging from 0.75 to 2.5 and $\alpha = 0.05$. For example, suppose that it is desired to detect a true mean difference of $\delta = 2$ kg of milk, and it is known from previous studies or experience at a particular commercial dairy that $\sigma_e^2 = 3 \text{ kg}^2$, $\sigma_{c(t)}^2 = 3 \text{ kg}^2$, and $\sigma_{p(t)}^2 = 1 \text{ kg}^2$. Suppose that this dairy groups animals in pens such that there are $d = 8$ animals per pen. Then the experimental variability between pens, being the experimental units, is defined by

$$\sigma_r^2 = \frac{\sigma_e^2 + \sigma_{c(t)}^2}{d} + \sigma_{p(t)}^2 = \frac{3 + 3}{8} + 1 = 1.75 \text{ kg}^2$$

such that the effect size is

Figure 1. Statistical power of test versus number of experimental units for effect sizes (Δ) equal to 0.75 ($\circ$), 1.00 ($\bullet$), 1.25 ($\square$), 1.50 ($\blacksquare$), 1.75 ($\diamond$), 2.00 ($\blacklozenge$), 2.25 ($\triangle$), or 2.50 ($\blacktriangle$) for comparing 2 treatments in a) completely randomized design, b) single crossover design, and c) randomized complete block design with subsampling.

$$\Delta = \frac{\delta}{\sigma_r} = \frac{2}{\sqrt{1.75}} \approx 1.5.$$

Using the $\Delta = 1.5$ curve from Figure 1a, $n = 4$ pens per treatment would provide less than 50% power, whereas $n = 8$ pens per treatment would be required for sufficient power (i.e., of 80%). Note that then $8 \times 8 = 64$ cows would be required per treatment.

Now suppose it was possible to have 20 instead of 8 animals per pen. Then using the same VC specifications as provided,

$$\sigma_r^2 = \frac{3+3}{20} + 1 = 1.2 \text{ kg}^2$$

(i.e., experimental variability decreases because of the larger pen size) such that

$$\Delta = \frac{2}{\sqrt{1.2}} \approx 1.75.$$

That is, nearly 6 pens per treatment would provide sufficient power using Figure 1a. Although the number of required pens is fewer with 20 cows as opposed to 8 cows per pen, the total number of cows per treatment would be greater (i.e., $6 \times 20 = 120$ cows per treatment).

It should be always clearly noted that $\sigma_r^2 \geq \sigma_{p(t)}^2$; that is, experimental variability will never be less than the variability that exists between pen effects, regardless of the number of animals per pen. That is, for the particular mean difference δ and VC specifications provided above, Δ is always less than 2, for instance,

$$\Delta < \frac{\delta}{\sqrt{\sigma_{p(t)}^2}} < \frac{2}{\sqrt{1}} = 2,$$

even if $d \rightarrow \infty$ (i.e., the number of cows in a pen increased indefinitely). For example, 4 pens per treatment would only ensure a maximum of 65% power using Figure 1a, regardless of d , the number of cows per pen. The only way to further increase power, if constrained by the use of 4 pens per treatment, would be to either decrease $\sigma_{p(t)}^2$ below 1 kg^2 , if the facilities could be modified accordingly (e.g., greater uniformity of pen environments), or be satisfied with a larger minimum true mean difference (i.e., $\delta > 2 \text{ kg}$) between the 2 treatments.

Now suppose that this study could be conducted such that animals are individually penned and fed (i.e., $d = 1$). Then, using the VC values as provided, $\sigma_r^2 = \sigma_e^2 + \sigma_{c(t)}^2 + \sigma_{p(t)}^2 = 7 \text{ kg}^2$ such that

$$\Delta = \frac{2 \text{ kg}}{\sqrt{7} \text{ kg}} = 0.760.$$

Hence, the effect size would be smaller for an individually treated animal study compared with a pen study, assuming that $\sigma_{p(t)}^2$ would be the same for individually fed versus group-fed animals. Using the $\Delta = 0.75$ curve from Figure 1a, 10 experimental units or cows per treatment would be insufficient for 40% power to detect $\delta = 2 \text{ kg}$; that is, generally fewer pens per treatment would be required for a pen study than cows per treatment for an individual cow study, as also pointed out by St-Pierre (2007). However, it should be noted that $\sigma_{p(t)}^2$ may be substantially smaller for individual animals in tie-stall facilities than for grouped animals in large commercial dairies, especially if pens in the latter are large or rather distal to each other. In spite of this possibility, one would likely need far fewer cows in total for individual cow studies than for pen studies. For example, based on the given VC specifications, 16 individually fed cows per treatment would provide 80% power to detect a mean difference of $\delta = 2.65 \text{ kg}$ ($\Delta = 1.00$), whereas 4 pens per treatment, each pen with 4 cows ($\Delta = 1.67$), would afford less than 55% power using Figure 1a. As a side note, $\sigma_{p(t)}^2$ and $\sigma_{c(t)}^2$ are not separately estimable in individual cow CRD studies such that their sum is used and relabeled as the between cow VC ($\sigma_{c(t)}^2$).

Single Crossover Design

Consider a 2-period simple crossover (i.e., 2×2 Latin square) study comparing the 2 treatments A and B involving n pens in which one-half ($n/2$) of the pens are assigned to the A→B treatment sequence with the remaining pens assigned to the B→A sequence over the 2 periods. With d animals per pen,

$$\sigma_r^2 = \frac{\sigma_e^{2*}}{d} + \sigma_{p^{*t}}^2$$

for this design. The 2 VC in the sum $\sigma_e^{2*} = \sigma_e^2 + \sigma_{c^{*t}}^2$ would be generally combined and labeled together as the within-cow VC, because its 2 components are not

separately estimable. Here, σ_r^2 is derived as the ANOVA EMS for pen $\times$ treatment, divided by the number (d) of observations on each experimental unit or pen $\times$ treatment combination. Note that, unlike the CRD, σ_r^2 no longer depends upon either the between-cow (σ_c^2) or between-pen variability (σ_p^2), because treatment comparisons are made both within cows and pens. However, σ_r^2 does depend upon the variability in true treatment differences across pens as measured by $\sigma_{p^{*t}}^2$. If $d = 1$, as in individual cow studies, then $\sigma_r^2 = \sigma_e^{2*} + \sigma_{p^{*t}}^2$; hence, it would not be further possible to separate any of the VC in σ_r^2 such that σ_r^2 would then be labeled as the within-cow VC. Nevertheless, as noted similarly for the CRD design, the magnitude of $\sigma_{p^{*t}}^2$ for individually treated animals may not be the same as that for group-treated animals because of the potentially different nature of pen effects for individually treated versus penned animals.

Statistical power versus n for the simple crossover design is provided in Figure 1b for various specifications of Δ ranging from 0.75 to 2.50. Note that n refers to the total number of pens used in the crossover study. Suppose that $\sigma_{p^{*t}}^2 = 0.5 \text{ kg}^2$, $\sigma_e^{2*} = 3.0 \text{ kg}^2$, and each pen is to have $d = 6$ cows. With an expected treatment mean difference of $\delta = 2 \text{ kg}$ and

$$\sigma_r^2 = \frac{\sigma_e^{2*}}{d} + \sigma_{p^{*t}}^2 = \frac{3}{6} + 0.5 = 1.0 \text{ kg}^2,$$

then

$$\Delta = \frac{2 \text{ kg}}{\sqrt{1} \text{ kg}} = 2.$$

Based on Figure 1b, a total of $n = 7$ pens, 3 pens on one sequence (i.e., A→B) and 4 on the other (i.e., B→A), would provide 80% power. Suppose, however, that $d = 22$ cows per pen are available, such that

$$\sigma_r^2 = \frac{3}{22} + 0.5 = 0.636 \text{ kg}^2$$

leading to

$$\Delta = \frac{2 \text{ kg}}{\sqrt{0.636} \text{ kg}} = 2.5.$$

From Figure 1b, it then should be apparent that $n = 6$ pens (3 per sequence) would now provide superior ($\sim 90\%$) power, whereas $n = 4$ pens (2 per sequence) would not. In fact, $\Delta = 4.0$ would be required to assure 80% power using $n = 4$ pens for the single crossover. However, this particular effect size could never be attained with the VC specifications provided here, even if $d \rightarrow \infty$. The limitation is that σ_r^2 always would be greater than $\sigma_{p^*t}^2$ such that the effect size for these particular specifications would have an upper bound of

$$\Delta < \frac{\delta}{\sqrt{\sigma_{p^*t}^2}} = \frac{2}{\sqrt{0.5}} = 2.83.$$

An effect size of $\Delta \geq 4$ could be achieved with $\sigma_{p^*t}^2 = 0.5$ if $\delta \geq 2.83$ kg, or it could be achieved with $\delta = 2$ kg if $\sigma_{p^*t}^2 \leq 0.25$ kg², or some combination of an increase in δ or a decrease in $\sigma_{p^*t}^2$ thereof. Please note that, regardless of the VC values, a minimum of 2 pens per treatment sequence (i.e., a total of $n = 4$ pens) would be absolutely necessary to conduct this design to estimate any potential period and period $\times$ treatment interaction effects in addition to treatment effects.

Randomized Complete Block Design with Subsampling

There has been another increasingly common misconception with studies on commercial dairies that if different treatments could be randomly assigned to cows within the same pen (e.g., using Calan gates), then cows would define the experimental units for comparing treatments. However, even then, it is still the number of pens that defines the number of experimental units. This would certainly be the case if all cows inadvertently received the same treatment within a pen. The experimental variability for this randomized complete block design with subsampling (RCBDS) design is inherently based on pen $\times$ treatment interaction, as also so clearly explicated by St-Pierre (2007).

With d animals assigned to each treatment within each pen (i.e., a pen size of 2 d animals), the experimental variability for the RCBDS is

$$\sigma_r^2 = \frac{\sigma_e^{2*}}{d} + \sigma_{p^*t}^2,$$

where $\sigma_e^{2*} = \sigma_e^2 + \sigma_c^2 + \sigma_{c^*t}^2$. As for the single crossover design, σ_r^2 is derived as the ANOVA EMS for pen $\times$

treatment, divided by the number (d) of observations on each experimental unit or pen $\times$ treatment combination. However, this specification of σ_r^2 is slightly larger with the RCBDS compared with the crossover design for 2 primary reasons: 1) σ_c^2 is now included, because treatments are not compared within cows as they were with the crossover design, and 2) d = number of animals per pen per treatment in the RCBDS, whereas d = number of cows per pen in the crossover design. That is, the pen density would be twice as great in the RCBDS compared with the crossover design for the same value of d .

Statistical power as a function of the number of pens (n) and effect size (Δ) for the RCBDS is presented in Figure 1c. From a quick comparison of Figures 1b and 1c, it might seem that for pen studies the single crossover design is more powerful compared with the RCBDS for the same value of Δ . This seems particularly so, because the size of σ_r^2 is larger with the RCBDS, thereby requiring a larger mean difference of δ for the RCBDS to compensate in determining Δ . Furthermore, the 2 different definitions of d reflect that greater pen densities would be required for the RCBDS to be competitive with the single crossover design. However, one practical disadvantage of the single crossover design is that the experiment needs to be conducted over 2 periods in contrast to the CRD or the RCBDS. Suppose that $\sigma_{p^*t}^2 = 0.5$ kg², whereas $\sigma_e^{2*} = 4$ kg² in the RCBDS, σ_e^{2*} being larger in the RCBDS than that specified for the single crossover example because of the additional $\sigma_c^2 = 1$ kg². Furthermore, suppose each treatment is to have $d = 6$ cows within each pen (i.e., pen size of 12 in the RCBDS). With an expected mean treatment difference of $\delta = 2$ kg and

$$\sigma_r^2 = \frac{\sigma_e^{2*}}{d} + \sigma_{p^*t}^2 = \frac{4}{6} + 0.5 = 1.17 \text{ kg}^2,$$

then

$$\Delta = \frac{2 \text{ kg}}{\sqrt{1.17 \text{ kg}}} = 1.85.$$

Interpolating between $\Delta = 1.75$ and the $\Delta = 2.00$ curves in Figure 1c, a total of $n = 7$ pens would be required to provide a statistical power close to 80%. Hence, this RCBDS study would require a total of 2 treatments $\times$ 7 pens/treatment $\times$ 6 cows/pen = 84 cows, whereas recall that the single crossover study required only 42

cows (7 pens of 6 cows each) to provide nearly the same power, albeit over 2 periods.

STATISTICAL POWER FOR BINARY RESPONSES

Many responses of interest involving research on commercial dairies involve reproduction or health characteristics that are measured on a binary scale (e.g., pregnant versus nonpregnant, diseased versus healthy). Generalized linear mixed models (**GLMM**) have been used by quantitative geneticists to provide genetic evaluations on animals for nonnormal or categorical responses such as calving ease in cattle (Gianola and Foulley, 1983). The probit or threshold mixed model is one such GLMM that is particularly well-suited and biologically interpretable for the analysis of various binary reproduction and health characters in dairy cattle (Tempelman, 1998). This model assumes that an observed binary outcome, for example, disease or pregnancy, is determined by an underlying and unobserved normally distributed variable called liability. If the liability for a certain animal exceeds a specific threshold, then the animal is observed to be diseased; otherwise, it is healthy (Falconer, 1986). For example, suppose that the binary outcome (yes or no) of clinical mastitis observed on any particular cow is determined by an underlying normally distributed disease load. If the disease load or liability of the animal exceeds a certain threshold, then a clinical mastitis outcome is observed; otherwise, it is not.

Labeling disease as the “success” outcome, the threshold mixed model is typically defined such that the probability of success for a treatment i is $p_i = \Phi(\mu_i)$, where $\Phi(\cdot)$ indicates the standard normal distribution function and μ_i indicates the mean response for treatment i on the underlying normal scale. Note that tabulated values of this function are commonly provided as the first appendix table near the back of most statistical methods textbooks. For example, $\Phi(1.96) = 0.975$ is well known for its use to provide 95% confidence intervals or 2-tailed z -tests on the means or mean differences. The inverse of this function is referred to as the probit function $\Phi^{-1}(\cdot)$; for instance, $\Phi^{-1}(0.975) = 1.96$.

Suppose, as before, one uses the CRD design whereby several pens are assigned at random to each treatment, each pen having d animals per pen. For an experimental unit or pen j assigned to a particular treatment i , the expected proportion of the d cows having the “success” outcome can be written using the GLMM as

$p_{ij} = \Phi(\mu_i + r_{ij})$, where r_{ij} = the random experimental effect associated with pen j within treatment i . Here, r_{ij} is a random draw from a normal distribution with variance

$$\sigma_r^2 = \frac{\sigma_e^2 + \sigma_{c(t)}^2}{d} + \sigma_{p(t)}^2$$

as presented earlier for the CRD. In other words, σ_r^2 models experimental variability, often referred to as overdispersion in GLMM (Tempelman, 1998), on the underlying liability scale.

Suppose that a binary health response is recorded and one wishes to know whether or not there is sufficient power to be able to reject $H_0: p_1 = p_2$ in favor of $H_1: p_1 \neq p_2$, where p_1 and p_2 represent the incidence rates of disease or pregnancy for 2 treatments (1 and 2), respectively. The mixed model power determination procedures of Stroup (2002) can be adapted for computing power for the threshold model. In a threshold model analysis using software such as SAS PROC GLIMMIX, the number of diseased cows over the total number of animals per each experimental unit (i.e., pen) defines the response variable (Littell et al., 2006); SAS PROC GLIMMIX will also estimate VC for any specified random effects. In a power analysis, one would substitute dp_i over d for the total number of diseased cows over all cows, respectively, for any pens associated with treatment i , specifying any VC contributing to σ_r^2 to be known or fixed. The resulting output F -test statistics can then be used to determine the noncentrality parameters that are subsequently used to approximate the power, similar to what is done in Stroup (2002) and Kononoff and Hanford (2006). Recently, Walt Stroup (University of Nebraska, Lincoln; personal communication) has determined based on simulation analyses that these power approximations are reliable for GLMM analysis of binary data.

Figure 2 has power curves to detect a difference between 2 incidence rates of $p_1 = 0.05$ and p_2 ranging from 0.10 to 0.45 for a CRD where cows are the experimental units (i.e., cows are individually treated). Furthermore, no overdispersion or experimental variability is assumed on the underlying scale, for instance, $\sigma_r^2 = 0$. In other words, the power determinations using Figure 2 are likely to be slightly overstated for real applications. Nevertheless, one can clearly see that at least 425 cows per treatment would be required to find a difference in incidence rates between 5 and 10%, whereas at least 150 cows per treatment would be required to find a

Figure 2. Statistical power of test for comparing 2 treatments for differences in incidence rates of binary responses given an incidence rate of 0.05 for 1 treatment and various incidence rates for the second treatment (0.10 ○, 0.15 ●, 0.20 □, 0.25 ■, 0.30 ◇, 0.35 ◆, 0.40 △, and 0.45 ▲) as a function of number of individually treated cows per treatment.

difference in incidence rates between 5 and 15%. Clearly, individual cow studies pose substantial logistical and cost difficulties when conducted on this scale.

Figure 3 displays a panel of power curves in a CRD to detect a difference between 2 incidence rates of $p_1 = 0.05$ and p_2 ranging from 0.10 to 0.45 and as a function of the number of pens per treatment for $d = 24$ or 48 cows per pen and

$$\sigma_r^2 = \frac{\sigma_e^2 + \sigma_{c(t)}^2}{d} + \sigma_{p(t)}^2 = 0.01, 0.10, \text{ or } 0.20.$$

Even when σ_r^2 is relatively small ($\sigma_r^2 = 0.01$), there would be no situation in which 2 pens per treatment with 24 cows per pen would provide sufficient power of test to detect a difference in incidence rates as large as that between $p_1 = 0.05$ and $p_2 = 0.45$ (Figure 3a). With 48 cows per pen, one would barely attain 80% power to detect a difference between $p_1 = 0.05$ and $p_2 = 0.45$ with 2 pens per treatment, even with low experimental variability ($\sigma_r^2 = 0.01$) between pens (Figure 3b). However, not even 4 pens per treatment with 48 cows per pen would be sufficient to detect a difference in incidence rates between 0.05 versus 0.15 with sufficient power. Of course, as σ_r^2 increases, power of test would further deteriorate as illustrated for $\sigma_r^2 = 0.10$ (Figure 3c and d) and $\sigma_r^2 = 0.20$ (Figure 3e and f). Neverthe-

less, pen studies have merit from a management viewpoint, because they require substantially fewer experimental units to achieve the same statistical power compared with large individually treated cow studies. For example, from Figure 3a, 5 pens per treatment, each pen with 24 cows, would be required to detect a difference in incidence rates between 0.05 and 0.20 with sufficient power, whereas from Figure 2, eighty-five individually treated cows per treatment would be required. Although the pen study requires more animals per treatment ($5 \times 24 = 120$) for the same power, it might be more convenient to manage 5 pens of animals per treatment than individually handle 85 cows per treatment.

The unfortunate limitation of Figure 3 is that it only applies to specifically chosen values of p_1 , p_2 , d , and σ_r^2 , whereas many readers may desire to consider alternative specifications. It is possible to derive an approximate effect size for a threshold mixed model analysis based on any combination of values of p_1 , p_2 , d , and σ_r^2 .

$$\Delta = \frac{|\Phi^{-1}(p_1) - \Phi^{-1}(p_2)|}{\sqrt{\frac{p_1(1-p_1)}{2d\varphi^2(\Phi^{-1}(p_1))} + \frac{p_2(1-p_2)}{2d\varphi^2(\Phi^{-1}(p_2))} + \sigma_r^2}}, \quad [1]$$

where

$$\varphi^2(x) = \left(\frac{1}{\sqrt{2\pi}} \exp(-0.5x^2) \right)^2$$

simply refers to the square of the standard normal probability density function evaluated at x . The numerator of Δ in equation [1] is a measure of the mean difference between the 2 treatments on the underlying liability scale, whereas the denominator of Δ defines a measure of variability that includes not only σ_r^2 , defining variability between experimental units, but also pure binomial variability. That is, even if there was no experimental variability between pens, such that the probability of having the success outcome was identical for all such pens, the resulting outcomes for different pens will still differ because of binomial sampling variability. Recall that σ_r^2 is a measure of overdispersion that takes into account additional biological or environmental sources of variability. Exploring equation [1] further, one could determine for the same difference between p_1 and p_2 that Δ , and hence statistical power, will be larger for detecting a particular difference be-

Figure 3. Statistical power of test for comparing 2 treatments for differences in incidence rates of binary responses given an incidence rate of 0.05 for 1 treatment and various incidence rates for the second treatment (0.10 $\circ$, 0.15 $\bullet$, 0.20 $\square$, 0.25 $\blacksquare$, 0.30 $\diamond$, 0.35 $\blacklozenge$, 0.40 $\triangle$, and 0.45 $\blacktriangle$) as a function of the number of pens per treatment. Experimental variability (σ_r^2) and number of cows (d) per pen set equal to a) $\sigma_r^2 = 0.01$, $d = 24$, b) $\sigma_r^2 = 0.01$, $d = 48$, c) $\sigma_r^2 = 0.10$, $d = 24$, d) $\sigma_r^2 = 0.10$, $d = 48$, e) $\sigma_r^2 = 0.20$, $d = 24$, and f) $\sigma_r^2 = 0.20$, $d = 48$.

tween 2 extreme proportions (e.g., between 0.05 and 0.10) than between 2 intermediate proportions (e.g., between 0.45 and 0.50).

Once Δ is computed using equation [1] for binary responses in the CRD design, then it can be used to determine power using Figure 1a. For example, from Figure 3d, a CRD with $n = 5$ pens per treatment, based on $d = 48$ cows per pen and an experimental variability between pens of $\sigma_r^2 = 0.10$ on the underlying liability scale, would provide nearly 80% power for detecting a difference between incidence rates of 0.05 and 0.20. Using these same specifications in equation [1], the numerator is

$$\left| \Phi^{-1}(0.05) - \Phi^{-1}(0.20) \right| = |-1.64 - -0.84| = 0.80.$$

The corresponding denominator is

$$\sqrt{\frac{0.05(1-0.05)}{2(48)\varphi^2(-1.64)} + \frac{0.20(1-0.20)}{2(48)\varphi^2(-0.84)}} + 0.10 = 0.41.$$

Hence, the corresponding value of Δ is nearly 2.00 ($\Delta = 1.95 = 0.80/0.41$). One could then use this value of Δ and $n = 5$ pens per treatment to draw the same power conclusion of roughly 80% from Figure 1a as we did from Figure 3d.

Similarly, one can use equation [1] for evaluating power for binary data for different single crossover scenarios in Figure 1b, remembering, however, that now n is equal to the total number of pens rather than the number of pens per treatment, and that the specification for σ_r^2 will also depend upon the study. For example, suppose that a single crossover design is used, and it is anticipated that the experimental variability is

$$\sigma_r^2 = \frac{\sigma_e^{2*}}{d} + \sigma_{p^*t}^2 = 0.10,$$

where $\sigma_e^{2*} = \sigma_e^2 + \sigma_{c^*t}^2$, recalling that d refers to the number of cows per pen used over the 2 periods. Also, suppose that $d = 48$ cows per pen are available. Using equation [1], this also translates to a $\Delta \approx 2.0$ such that, based on Figure 1b, 7 pens would be needed to provide 80% power, thereby requiring fewer cows (7 pens $\times$ 48 cows/pen = 336 cows) than the CRD (2 treatments $\times$ 5 pens/treatment $\times$ 48 cows/pen = 480 cows) for the same power. However, as previously noted, the crossover experiment would need to be conducted over 2

periods. Finally, note that equation [1] could be similarly used in conjunction with Figure 1c to determine the power for the RCBDS when the responses are binary.

ELICITING VC SPECIFICATIONS

One often troubling question is what are appropriate VC values to specify for a statistical power analysis for pen studies? On many commercial dairies, it should be possible to use historical records for some traits such as milk production to have some indication of the magnitude of the various VC. Furthermore, based on extensive quantitative genetics research using test-day models, reasonable specifications for σ_e^2 vary from 2 kg² in late lactation to 10 kg² in early lactation, with σ_e^2 generally increasing with parity (Jensen, 2001; Tempelman, 2004). Given that the reported repeatabilities $r = \sigma_c^2 / (\sigma_c^2 + \sigma_e^2)$ (Van Vleck, 1993) for milk production often approach 0.50 (Dong et al., 1988), this also suggests that σ_c^2 is of a similar magnitude as σ_e^2 . However, it seems that values of σ_c^2 and σ_e^2 might be herd-dependent (Weigel et al., 1993), and it is even more likely that values of σ_p^2 might vary across herds for pen studies. In other words, there is likely greater variability between pen performance within some herds than in others due to differences in management or orientation of pens.

Because not enough dairy studies may have been conducted with the necessary precision to determine suitable values for various VC of interest, one reasonable strategy for eliciting VC specifications is based on the commonly used range approximation as expounded in various statistical methods textbooks (Gill, 1978; Ott and Longnecker, 2001). The empirical rule for the normal distribution dictates that 95% of the effects for a particular random factor (e.g., cows or pens or herds) should fall within ± 2 standard deviations of the mean, or equivalently, within a range of 4 standard deviations. Here, standard deviations represent the square root of the VC of interest. Hence, the range (R) of the effects divided by 4 roughly approximates the standard deviation or square root of the respective VC for the random effects in question. For example, suppose that on a particular farm, one expects the range of true mean pen performances for milk production (on the same treatment) to be as large as 2 kg between pens of greatest and least production. Then, $\sigma_p \approx R / 4 = 2 / 4 = 0.5$ kg such that $\sigma_p^2 \approx 0.25$ kg² would be a reasonable specification for that VC.

One could also elicit VC for random interaction effects involving treatments like $\sigma_{p^*t}^2$ or $\sigma_{c^*t}^2$ in a similar manner. Suppose that R_δ denotes the range of differences between 2 treatments across levels of the random effects factor (e.g., pens); then,

$$\frac{R_\delta / 4}{\sqrt{2}}$$

defines the square root of the corresponding VC (e.g., $\sigma_{p^*t}^2$). For example, suppose that the true mean difference between 2 treatments is specified to be $\delta = +2$ kg in a RCBDs but that one anticipated the range of true mean differences between 2 treatments to be pen-dependent, varying from +1 to +3 kg across pens, thereby leading to a range of $R_\delta = 2$. Then, the range approximation as it applies to these differences would be such that

$$\sigma_{p^*t} \approx \frac{R_\delta / 4}{\sqrt{2}} = \frac{2 / 4}{\sqrt{2}} = \frac{\sqrt{2}}{4}$$

or

$$\sigma_{p^*t}^2 \approx \left(\frac{\sqrt{2}}{4} \right)^2 = 0.125 \text{ kg}^2.$$

Eliciting appropriate values for VC in threshold mixed models is a little more challenging, because that variability is characterized on an underlying unobserved scale. Furthermore, the observed effect of a particular VC value at one incidence rate is not necessarily the same as it is for another incidence rate. Extending the empirical rule, Figure 4 demonstrates the magnitude of $4\sigma_r$ on the true range in incidence rates across experimental units by plotting 3 sets of effective range bands (corresponding to 3 different overall incidence rates: $p = 0.05$, $p = 0.25$, and $p = 0.50$) such that the width of the bands are defined by

$$\left[\Phi\left(\Phi^{-1}(p) - 2\sigma_r\right), \Phi\left(\Phi^{-1}(p) + 2\sigma_r\right) \right]$$

as a function of σ_r^2 . For example, the effect of an experimental variability of $\sigma_r^2 = 0.05$ on the underlying liability scale leads to an effective range of pen effects in incidence rates of [0.02, 0.12] when $p = 0.05$, of [0.13,

Figure 4. Empirical ranges (curves) of observed incidence rates of binary responses between experimental units (e.g., pens) based on overall incidence rates of $p = 0.05$ (solid line), $p = 0.25$ (short-dash line), and $p = 0.50$ (long dash line) as a function of the experimental variability (σ_r^2) on the underlying liability scale.

0.41] when $p = 0.25$, and of [0.33, 0.67] when $p = 0.50$. That is, greater variability in true incidence rates is expected between different experimental units when overall incidence rates are intermediate and when σ_r^2 is large. So, for instance, if one anticipated a range in true incidence rates between experimental units (i.e., pens) on the same treatment to be as great as [0.13, 0.41], then $\sigma_r^2 = 0.05$ would be a reasonable specification.

SCOPE OF INFERENCE

Thus far, this paper has concentrated entirely on ordinary replications; that is, the replication that ensures that treatments are replicated across several experimental units within a particular herd, whether they be pens or individually treated animals. However, if the intent of the study is that inferences from the study are applied to all herds, from which the study herd is considered to be somewhat representative, then herd effects are considered to be random effects. Hence, a single herd study provides no replication. That is, the scope of inference for a single herd study is such that any conclusions derived from such a study would strictly only apply to that particular herd. Of course, the same criticism could be levied against university-based experiment station research such that there should be continued support for multistate or regional university projects that involve replicating experiments at several experiment stations (Stevenson et al., 2006). Nevertheless, research station effects should be specified

as random in multiuniversity projects such that treatment $\times$ station interaction serves as the error term for testing treatment effects; this would effectively specify research station as the experimental unit, much like what is done for meta-analysis when studies are treated as experimental units (St-Pierre, 2001).

If an on-farm study involves multiple herds, such that each treatment is used within each herd, and it is desired that the scope of inference includes all herds from which those used in the study are a random sample, then the treatment $\times$ herd interaction serves as the experimental error term for treatment. This specification is analogous to a RCBDS except that now each pen within a herd is considered to be the subsample. Therefore, for the situation in which a CRD is considered within each of several herds, the experimental variability to allow the broader scope of study should be modified to be

$$\sigma_r^{2*} = \frac{\sigma_r^2}{b} + \sigma_{h*t}^2,$$

where σ_{h*t}^2 = the VC due to herd $\times$ treatment interaction as previously noted; b = the number of pens per treatment per herd; and

$$\sigma_r^2 = \frac{\sigma_e^2 + \sigma_{c(t)}^2}{d} + \sigma_{p(t)}^2$$

as specified previously for the CRD but now with all terms nested within herds. Then one could use Figure 1b to determine power for various values of

$$\Delta = \frac{\delta}{\sqrt{\sigma_r^{2*}}}$$

as based on the number of experimental units, n , now defined by the number of herds rather than the number of pens. The experimental variability for the other 2 designs discussed in this paper, the single crossover and RCBDS, could be similarly modified to provide the broader scope of inference with a multiple-herd study. Note that if herds are treated as fixed effects in a multi-herd study, there will be concomitant benefits in power but with the detriment that the results so obtained only apply to those specific herds (i.e., the inference space is narrowed; McLean et al., 1991).

The importance of this level of metareplication across herds is further borne out in a more general discussion by Johnson (2006). Metareplication within a somewhat

different context, involving the joint or meta-analysis of results from different dairy research studies, is discussed further by St-Pierre (2001), who also noted that each study (i.e., herd) should be treated as an experimental unit. Although the scope of inference is much greater with a multiherd study, far more resources are obviously required. Note that the experimental variability σ_r^{2*} will be affected primarily by either σ_h^2 or σ_{h*t}^2 , which represent variability between herds and their interactions with treatment, respectively. Hence, it would be vitally important to model potentially important herd-specific covariates that might lower either or both VC to increase statistical power. For example, it is well known that ambient temperature and humidity (Ravagnolo and Misztal, 2002) and herd size (Oleggini et al., 2001) influence various measures of dairy performance such that modeling their effects and their interactions with treatments could effectively decrease σ_h^2 or σ_{h*t}^2 , respectively. Consequently, σ_r^{2*} would be lowered such that Δ and hence statistical power would be enhanced.

Given the currently high inbreeding rates with dairy cattle (Weigel, 2001) and the concomitant increase in genetic relationships between animals both within and across herds, it might seem useful to consider modeling genetic effects to account for additional variability both within and between pens or herds. In fact, SAS PROC MIXED can be modified to do this provided one has a pedigree file available on all animals in the study (Tempelman and Rosa, 2004). Because of the complexity of the covariance structure in this case, it would be important to use the Satterthwaite or Kenward-Rogers option for estimating denominator degrees of freedom, as available in SAS PROC MIXED/GLIMMIX, for testing treatment differences (Littell et al., 2006). As multiherd databases accumulate and SNP genetic markers become progressively cheaper for cattle (Van Tassell et al., 2008), there will be more opportunities to pursue on-farm research that tailors drugs or feedstuffs for particular cow genotypes (Swanson, 2008), thereby providing some direction toward potentially more optimal strategies for grouping dairy cattle in pens. The scale of SNP association studies would likely require the need for thousands of animals (Iles, 2008) and hence the need for participating commercial dairies.

The use of multivariate mixed model analysis (Thompson and Meyer, 1986) would also facilitate the use of multiherd studies in which several performance traits (e.g., milk yield and its various components) are highly correlated. The benefits of greater precision on treatment effects for any one trait within a multivariate analysis over a series of trait-specific analyses was re-

cently reviewed by Riley et al. (2007). The advantages of multivariate analysis are particularly noticeable when measurements on some traits are missing or not extensively recorded.

There are some complications associated with pen studies on commercial dairies that go beyond the issue of scope of inference and the definition of the experimental unit. One is the issue of commingling dairy cows from different groups, particularly if it happens on a regular (i.e., weekly) basis. Some commingling might be accommodated with the RCBDS as treatments are compared within pens. However, it is difficult to know how to accommodate this mixing with a CRD design. If mixing is common, it would be useful to measure a response that can be recorded as regularly as the mixing itself (e.g., daily rather than monthly milk yields) such that adjustments for important covariates (e.g., DIM) can be modeled using polynomial (Jensen, 2001) or spline (White et al., 1999) relationships. Another related complication relates to the allelomimetic nature of dairy cattle behavior (Nordlund and Cook, 2006) whereby, for example, cows are stimulated to eat when other cows are eating, such that pen feeding might not be entirely representative of individual feeding for its biological effect. However, this phenomenon might be somewhat negated by competition effects due to antagonistic interactions, particularly in overstocked pens, which may further complicate the proper specification of between-pen and between-cow VC (Van Vleck and Cassady, 2005). Inferences then on effects of interest (e.g., parity $\times$ treatment interaction) may be biased or too complex to estimate if animals of different parities are grouped together, realizing that multiparous animals are generally socially dominant to primiparous animals (Nordlund and Cook, 2006). Finally, some dairies may have only 1 group per production level or stage of lactation, further necessitating the use of a multiherd study, given the inability to replicate within herds.

CONCLUSIONS

Based on power analyses presented in this paper, it is difficult to imagine any scenario involving the comparison of 2 treatments in which 4 pens would provide enough replication, even when treatments are compared within pens as they are in the single crossover or the RCBDS. In fact, 4 pens would be minimally necessary just to estimate experimental error in the single crossover design. Naturally, an even greater number of pens would be required for comparing more than 2 treatments. Given that it might be particularly difficult to attain more than 1 pen per treatment per dairy, and that replication across herds provide a scope of inference that is broader than what is possible with a single

dairy study, multiherd studies should be encouraged. These studies would then provide a scope of inference that is broader than what is possible with a single herd or single experiment station study.

REFERENCES

- Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Erlbaum, Hillsdale, NJ.
- Cox, D. F. 1980. Design and analysis in nutritional and physiological experimentation. *J. Dairy Sci.* 63:313–321.
- Dong, M. C., L. D. Van Vleck, and G. R. Wiggins. 1988. Effect of relationships on estimation of variance components with an animal model and restricted maximum likelihood. *J. Dairy Sci.* 71:3047–3052.
- Falconer, D. S. 1986. *Introduction to Quantitative Genetics*. 3rd ed. Longman Scientific and Technical, New York, NY.
- Gianola, D., and J. L. Foulley. 1983. Sire evaluation for ordered categorical data with a threshold model. *Genet. Sel. Evol.* 15:201–223.
- Gill, J. L. 1978. *Design and Analysis of Experiments in the Animal and Medical Sciences*. Iowa State University Press, Ames.
- Goldberger, A. S. 1962. Best linear unbiased prediction in the generalized linear regression model. *J. Am. Stat. Assoc.* 57:369–375.
- Henderson, C. R. 1953. Estimation of variance and covariance components. *Biometrics* 9:226–252.
- Henderson, C. R., O. Kempthorne, S. R. Searle, and C. M. Vonkrosigk. 1959. The estimation of environmental and genetic trends from records subject to culling. *Biometrics* 15:192–218.
- Iles, M. M. 2008. What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet.* 4:e33.
- Jensen, J. 2001. Genetic evaluation of dairy cattle using test-day models. *J. Dairy Sci.* 84:2803–2812.
- Johnson, D. H. 2006. The many faces of replication. *Crop Sci.* 46:2486–2491.
- Kononoff, P. J., and K. J. Hanford. 2006. Estimating statistical power of mixed models used in dairy nutrition experiments. *J. Dairy Sci.* 89:3968–3971.
- Littell, R. C., G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger. 2006. *SAS for Mixed Models*. 2nd ed. SAS Institute Inc., Cary, NC.
- McLean, R. A., W. L. Sanders, and W. W. Stroup. 1991. A unified approach to mixed linear models. *Am. Stat.* 45:54–64.
- Milliken, G. A., and D. E. Johnson. 1984. *Analysis of Messy Data. Volume I: Designed Experiments*. Wadsworth Inc., Belmont, CA.
- Nordlund, K. V., and N. B. Cook. 2006. Commingling dairy cows: Pen moves, stocking density, and health. Pages 36–42 in *The American Association of Bovine Practitioners Proceedings*. Frontier Printers, Stillwater, OK.
- O'Brien, R. G., and V. I. Lohr. 1984. Power analysis for univariate linear models: The SAS system makes it easy. Pages 840–846 in *Proceedings of the Ninth Annual Conference*, Cary, NC. SAS Users' Group International, SAS Institute Inc., Cary, NC.
- Oleggini, G. H., L. O. Ely, and J. W. Smith. 2001. Effect of region and herd size on dairy herd performance parameters. *J. Dairy Sci.* 84:1044–1050.
- Ott, R. L., and M. Longnecker. 2001. *An Introduction to Statistical Methods and Data Analysis*. 5th ed. Duxbury, Pacific Grove, CA.
- Ravagnolo, O., and I. Misztal. 2002. Effect of heat stress on nonreturn rate in Holsteins: Fixed-model analyses. *J. Dairy Sci.* 85:3101–3106.
- Riley, R. D., K. R. Abrams, P. C. Lambert, A. J. Sutton, and J. R. Thompson. 2007. An evaluation of bivariate random-effects meta-analysis for the joint synthesis of two correlated outcomes. *Stat. Med.* 26:78–97.
- St-Pierre, N. R. 2001. Invited review: Integrating quantitative findings from multiple studies using mixed model methodology. *J. Dairy Sci.* 84:741–755.

- St-Pierre, N. R. 2007. Design and analysis of pen studies in the animal sciences. *J. Dairy Sci.* 90:E87–E99.
- St-Pierre, N. R., and L. R. Jones. 1999. Interpretation and design of nonregulatory on-farm feeding trials. *J. Anim. Sci.* 77:177–182.
- Stevenson, J. S., J. R. Pursley, H. A. Garverick, P. M. Fricke, D. J. Kesler, J. S. Ottobre, and M. C. Wiltbank. 2006. Treatment of cycling and noncycling lactating dairy cows with progesterone during ovsynch. *J. Dairy Sci.* 89:2567–2578.
- Stroup, W. W. 2002. Power analysis based on spatial effects mixed models: A tool for comparing design and analysis strategies in the presence of spatial variability. *J. Agric. Biol. Environ. Stat.* 7:491–511.
- Swanson, K. S. 2008. Using genomic biology to study liver metabolism. *J. Anim. Physiol. Anim. Nutr. (Berl.)* 92:246–252.
- Tempelman, R. J. 1998. Generalized linear mixed models in dairy cattle breeding. *J. Dairy Sci.* 81:1428–1444.
- Tempelman, R. J. 2004. Experimental design and statistical methods for classical and bioequivalence hypothesis testing with an application to dairy nutrition studies. *J. Anim. Sci.* 82:E162–E172.
- Tempelman, R. J., and G. J. M. Rosa. 2004. Empirical Bayes approaches to mixed model inference in quantitative genetics. Pages 149–176 in *Genetic Analysis of Complex Traits Using Sas*. A. M. Saxton, ed. SAS Institute Inc., Cary, NC.
- Thompson, R., and K. Meyer. 1986. A review of theoretical aspects in the estimation of breeding values for multi-trait selection. *Livest. Prod. Sci.* 15:299–313.
- Van Tassell, C. P., T. P. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods* 5:247–252.
- Van Vleck, L. D. 1993. *Selection Index and Introduction to Mixed Model Methods*. CRC Press Inc., Boca Raton, FL.
- Van Vleck, L. D., and J. P. Cassady. 2005. Unexpected estimates of variance components with a true model containing genetic competition effects. *J. Anim. Sci.* 83:68–74.
- Weigel, K. A. 2001. Controlling inbreeding in modern breeding programs. *J. Dairy Sci.* 84:E177–E184.
- Weigel, K. A., D. Gianola, B. S. Yandell, and J. F. Keown. 1993. Identification of factors causing heterogeneous within-herd variance components using a structural model for variances. *J. Dairy Sci.* 76:1466–1478.
- White, I. M. S., R. Thompson, and S. Brotherstone. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. *J. Dairy Sci.* 82:632–638.