J. Dairy Sci. 99:1–9
http://dx.doi.org/10.3168/jds.2016-11516

Short communication: On recognizing the proper experimental unit in animal studies in the dairy sciences

Nora M. Bello,*[1] Matthew Kramer,† Robert J. Tempelman,‡ Walter W. Stroup,§ Normand R. St-Pierre,# Bruce A. Craig,‖ Linda J. Young,¶ and Edward E. Gbur**
*Department of Statistics, Kansas State University, Manhattan 66506
†USDA, Agricultural Research Service, Beltsville, MD 20705
‡Department of Animal Science, Michigan State University, East Lansing 48824
§Department of Statistics, University of Nebraska, Lincoln 68583
#Department of Animal Sciences, The Ohio State University, Columbus 43210
‖Department of Statistics, Purdue University, West Lafayette, IN 47907
¶National Agricultural Statistics Service, Washington, DC 20250
**Agricultural Statistics Laboratory, University of Arkansas, Fayetteville 72701

ABSTRACT

Sound design of experiments combined with proper implementation of appropriate statistical methods for data analysis are critical for producing meaningful scientific results that are both replicable and reproducible. This communication addresses specific aspects of design and analysis of experiments relevant to the dairy sciences and, in so doing, responds to recent concerns raised in a letter to the editor of the *Journal of Dairy Science* regarding journal policy for research publications on pen-based animal studies. We further elaborate on points raised, rectify interpretation of important concepts, and show how aspects of statistical inference and elicitation of research conclusions are affected.
Key words: experimental unit, replication, observational unit, hierarchical data structure, pen

Short Communication

Sound design of experiments and proper implementation of appropriate statistical methods for data analysis are critical for producing meaningful scientific results that are both replicable and reproducible (Milliken and Johnson, 2009). First, consider the concept of a "statistical unit," as proposed by Robinson (2016) in a recent Letter to the Editor in the *Journal of Dairy Science*, a term that is decidedly vague and lacks a universal definition in the mainstream design of experiments literature, particularly for agricultural applications (Kuehl, 2000; Littell et al., 2006; Casella, 2008; Milliken and Johnson, 2009; Stroup, 2013). Instead, let us define the "experimental unit" and the "observational unit," both formally and in the specific context of the dairy sciences. The leading literature in design of experiments defines the experimental unit, also called the unit of replication, as *the smallest entity that is assigned independently of all other units to a particular treatment*; the word *independent* is key to this definition (Kuehl, 2000; Littell et al., 2006; Casella, 2008; Milliken and Johnson, 2009; Stroup, 2013). Experimental units are often assumed to be "exchangeable," a statistical term that implies that the units do not differ in any fundamental way, so that reliable inferences would be obtained regardless of which treatment was assigned to each unit.

In the dairy sciences, individual cows can sometimes serve as experimental units; for example, if treatments were different types or doses of antibiotics individually injected to treat mastitis. Even then, cows may still be housed together in pens but individual cows within a pen are randomly assigned to different treatments. In dairy nutrition, it is often of interest to compare diets that, for logistical reasons, are commonly fed (i.e., randomly assigned) to pens, such that all cows in the same pen are offered the same diet. For example, if one wanted to compare 2 diets, one could design an experiment by randomly assigning diets A and B each to a different random set of pens, with each pen holding several cows. In this case, the pen is clearly the experimental unit. If 2 pens receiving different diets showed any difference in outcome, we would not know whether this difference was due to the intended diet effect, a confounded pen effect, or a combination of both effects. To effectively separate diet effects from pen effects would require more pens; that is, diets need to be replicated to multiple pens. How many more pens? This is a question of statistical power and depends on how large the diet effect is expected to be, how variable observations from pens fed the same diet are, and how

Received May 26, 2016.
Accepted July 31, 2016.
[1]Corresponding author: nbello@ksu.edu

this variability partitions into pen-level (i.e., between-pen) variability and cow-level (i.e., within-pen) variability. For further details on statistical power in the context of the dairy sciences, the reader may refer to Tempelman (2009).

Distinct from an experimental unit, to which a treatment is independently applied, is the concept of an *observational unit*, also known as the sampling unit. This distinction is recognized in the response to Robinson (2016) by Lamberson (2016). An observational unit is defined as *the physical entity on which an outcome of interest is measured in an experiment* (Kuehl, 2000; Casella, 2008). In many simple designs, experimental units and observational units are synonymous; that is, they can be matched to the same physical entity (Kuehl, 2000; Littell et al., 2006; Stroup, 2013). This was true in the prior example when assessing the effect of antibiotic treatments individually injected and can also be true for the diet example if the outcome of interest were measured at the pen level (e.g., total intake for the pen or total time spent feeding for all animals in a pen). If pen is the entity that is both independently assigned to treatment and measured for outcome, then pen serves as both the experimental unit and the observational unit. On the other hand, if the outcome of interest in the diet example was measured on individual cows in each pen, say milk yield, one encounters a natural "gap" or "mismatch" between the entity independently assigned to treatment (i.e., pen) and the entity measured (i.e., individual cow within a pen). This is an example of a nested design structure: the pen is nested within a treatment and the individual cow is nested within a pen, thereby creating a hierarchical structure in the data.

A hierarchical data structure refers to a configuration of the data where observations are not mutually independent but rather have a correlation structure imposed by the experimental design. In our dairy example with diets applied to pens, pens consist of individual cows but these animals are not mutually independent and, consequently, neither are their observations. Specific biological reasons to explain lack of independence of observations collected on cows within a pen are context specific. In the dairy sciences, one can often anticipate within-pen dynamics; for instance, differential feed access due to social behavior (e.g., dominance) or management practices (i.e., feed mixing). Notably, this kind of correlation between observations from cows within a pen is different from a general "pen" effect, which may be due, for instance, to pen size, condition of the substrate, or shade availability, to name a few. It is precisely due to this correlation (i.e., lack of independence) between cow-level observations that it is not possible to separate diet effects from pen effects in a nutrition study conducted on only 2 pens, regardless of the number of cows in each pen. Whenever observational units are nested within an experimental unit, as is the case here, the observational units are commonly referred to as subsamples, pseudoreplicates, or technical replicates (Casella, 2008) to indicate that these observations are correlated and thus do not constitute true independent replication. Data structures such as these are common in the animal sciences; examples include multi-farm studies, groups of animals entering a study in weekly clusters, or repeated observations collected over time on individual animals (i.e., test-day milk yield). Hierarchical data structures, and thus underlying correlations between observations, can often be recognized as nesting or blocking in the experimental design of a study. Both nesting and blocking are common elements of design in dairy trials; thus, it is not surprising that experimental units are often separate physical entities from observational units in dairy science experiments.

We emphasize: *experimental units are defined in terms of independent treatment assignments whereas observational units are defined in terms of outcome measurements*. These are clearly different definition criteria. As such, observations do not necessarily represent replications. However, observational units are usually contained within experimental units (Stroup, 2013), which in turn determine the amount of replication of a given experiment. As a side note, a potential exception is a repeated-measures design, and this depends on whether one labels the observational unit to be an individual cow or an individual cow at a specific time point—here, labels are less important than the concept that repeated measures on the same cow are mutually correlated. Even so, replication implies an independent repetition of a basic experimental component, such as a treatment, and is considered a prime requisite for valid and reliable experimental inference (Kuehl, 2000; Casella, 2008). The rationale to support true replication as a requisite for valid experiments is well explained by Kuehl (2000), including the following: (1) results are reproducible, at least under the specified experimental conditions; (2) results are not aberrant realizations of an experiment due to unforeseen circumstances; and (3) variability between experimental units defining experimental error is properly estimated and thus subsequent hypothesis tests are properly calibrated.

To be able to identify hierarchical data structure; that is, when independent replication occurs and when it does not, it is most important to understand the complete process involved in collecting data and carrying out a study. This understanding is also critical to adequately specify the statistical model for data analysis. For illustration purposes, consider alternative layouts for a general 3×3 Latin square design consisting of 3

treatments, 3 periods, and 3 pens, with multiple cows per pen, thereby responding directly to the cases proposed by Robinson (2016). It should be noted that the case presented in Robinson's letter to the editor lacks a clear description of how the experiment was conducted, neither was the process of data collection clearly explained. As a result, the reader may surmise 2 plausible experimental scenarios, each leading unambiguously to 1 of his 2 models. First, suppose a scenario A, in which dietary treatments are fed to cows via a common pen trough (i.e., treatment is randomly assigned to pen). For contrast, we also consider a scenario B, whereby treatments are randomly assigned and applied to individual cows within a pen (e.g., via injection or feeding through Calan gate technology). For both scenarios, let us work through the exercise called "What Would Fisher Do?" (**WWFD**), which was introduced by Stroup (2013) to translate the description of an experimental design to an ANOVA shell to an actual statistical model. The WWFD exercise is a general strategy that, in the context of data assumed to be normal, can be shown to be equivalent to the traditional mean-squares ANOVA exercise (Milliken and Johnson, 2009) essential to identifying the proper experimental error and thus distinguish between an experimental unit and an observational unit. Figures 1 and 2 depict the implementation of the WWFD exercise to alternative layouts of a single (i.e., unreplicated) 3 × 3 Latin square design consisting of 3 treatments, 3 periods, and 3 pens, with multiple cows per pen. To follow the WWFD approach, it is important to note the relative positions of rows corresponding to treatment structure (i.e., the central column of the WWFD table) and rows corresponding to elements of the experimental design (i.e., left-most column of the WWFD table), as well as their combination (i.e., right-most column of the WWFD table) to properly characterize the data collection process. For technical details on the WWFD exercise, the interested reader is referred to Stroup (2013) and Stroup (2015). Figure 1 illustrates the WWFD exercise implemented for scenario A: treatments (e.g., diets) randomly assigned to pens. Here, random assignment of treatment is to pen within a given period, with reassignment of treatments at the beginning of each new period. Hence, pen within a period is the unit of randomization, and thus the experimental unit for treatment; this is reflected in the position of the term "pen × period" immediately below the term "treatment" in the left-most and central columns of the WWFD table, respectively (Figure 1). In the "combined" section of the WWFD table, "pen × period | treatment" is read "pen × period after accounting for treatment"; its degrees of freedom are specified by subtracting the "treatment" degrees of freedom from those of the "pen × period" element of design. Being the experimental unit for treatment, the pen in a given period defines the level of independent replication for treatment in the hierarchical data structure and thus identifies the experimental error term. This specification of pen in a given period as the experimental unit for treatment is neither optional nor subject to opinion—it is the way the experiment was set up. In turn, cow within a pen in a given period represents the observational unit, such that the term "period × cow(pen)" in the WWFD table defines the sampling error but not the experimental error (Figure 1). For completeness, we note that the sampling error term represents variation among observational units, distinct from experimental error or variation among experimental units. Furthermore, recall that for a single (i.e., nonreplicated) Latin square design, the interaction between treatment and pen, as well as that between treatment and period, are assumed nonexistent to allow for estimation of a measure of error (Kuehl, 2000; Milliken and Johnson, 2009) and are thus not considered in Figure 1. Adding the term "cow(pen)" in the WWFD table to recognize multiple cows measured in a pen does not override this assumption.

For contrast, consider the WWFD exercise implemented for scenario B: treatments assigned to individual cows within a pen (e.g., individual antimicrobial injections), as illustrated in Figure 2. The actual source terms in the WWFD table are similar to those shown for scenario A in Figure 1, but their relative positions in the table are modified to reflect differences in the randomization process and in data collection. More specifically, scenario B differs from scenario A in the relative position of the "treatment" row relative to the rows of elements of the experimental design of the WWFD table (Figures 1 and 2). In scenario B, treatments are randomly assigned to individual cows within a pen in a given period, with reassignment of treatments at the beginning of each new period. Hence, "period × cow(pen)" identifies the individual cow in a given period and constitutes the unit of randomization, and thus the experimental unit (Figure 2). In this scenario, the individual cow in a given period is also the observational unit on which the outcome is measured, as identified by the bottom row in the WWFD table (Figure 2). In turn, the "pen × period" term in scenario B identifies the pen within a period as an effective blocking structure within which treatments are randomly allocated to individual cows in pens.

In the dairy sciences, experiments with a Latin square design are sometimes repeated with more than one square, yielding so-called replicated Latin squares or Latin rectangles (Kuehl, 2000). In this case, one can— and should—investigate the interaction between treatment and period, provided the same periods are

Experimental design or design structure		Treatment structure		Combined	
Source	df	Source	df	Source	df
Pen	3 – 1 = 2			Pen	2
Cow(Pen)	3 × (125 – 1) = 372			Cow(Pen)	372
Period	3 – 1 = 2			Period	2
		Treatment	3 – 1 = 2	Treatment	2
Pen × Period	2 × 2 = 4			Pen × Period | Treatment	4 – 2 = 2
Period × Cow(Pen)	2 × 3 × (125 – 1) = 744			Period × Cow(Pen)	744
Total					1,124

Figure 1. "What Would Fisher Do?" exercise for scenario A: treatments independently assigned to pens within a period, based on Stroup (2013, 2015). This experiment has the general design structure of a single (i.e., unreplicated) 3 × 3 Latin square, consisting of 3 treatments, 3 periods, and 3 pens of 125 cows each. The arrow indicates the position of the row corresponding to treatment structure relative to a row of the experimental design that identifies the unit of randomization for treatment, and thus, its experimental unit. For replicated Latin squares, the interaction between period and treatment should also be considered (Tempelman, 2004).

considered within each square, particularly if period were reflective of a physiological event (e.g., days in milk or time since calving; Tempelman, 2004). Further, notice that this far, we have treated period as an element of experimental design (i.e., left-most column of the WWFD table in Figures 1 and 2), as is consistent with the general literature on design of experiments. However, specifically for some dairy applications, period may be legitimately considered either as an element of the experimental design or as an element of treatment structure, depending on the variability of days in milk within a period, and thus the stage of lactation (refer to Tempelman, 2004 for further details). The decisions on how to treat period (i.e., as an element of experimental design or as one of treatment structure), as well as the incorporation of the treatment × period interaction in replicated Latin squares, are not trivial and need to be made on a case-by-case basis. Indeed, studies based on replicated Latin square designs showed that dietary effects may depend on stage of lactation (Taylor and Allen, 2005), in which case inference should focus on treatment differences within periods (presuming period as an element of the treatment structure) as opposed to overall treatment effects. Further, treating period as an element of experimental design or of treatment structure has implications for downstream inference because it determines how the experimental error for the treatment of interest is defined (Tempelman, 2004).

Experimental design Pretreatment		Treatment structure		Combined	
Source	df	Source	df	Source	df
Pen	3 – 1 = 2			Pen	2
Cow(Pen)	3 × (125 – 1) = 372			Cow(Pen)	372
Period	3 – 1 = 2			Period	2
Pen × Period	2 × 2 = 4			Pen × Period	4
		Treatment	3 – 1 = 2	Treatment	2
Period × Cow(Pen)	2 × 3 × (125 – 1) = 744			Period × Cow(Pen)| Treatment	744 – 2 = 742
Total					1,124

Figure 2. "What Would Fisher Do?" exercise for scenario B: treatments independently assigned to individual cows within a pen in a given combination, based on Stroup (2013, 2015). This experiment has the general design structure of a single (i.e., unreplicated) 3 × 3 Latin square consisting of 3 treatments, 3 periods, and 3 pens of 125 cows each, but assigns treatments within each pen-by-period combination. The arrow indicates the position of the row corresponding to treatment structure relative to a row of the experimental design that identifies the unit of randomization for treatment, and thus, its experimental unit. For replicated Latin squares, the interaction between period and treatment should also be considered (Tempelman, 2004).

Once the WWFD exercise has been completed so that the randomization and data collection processes in scenarios A and B are fully characterized, one can then transfer the row elements of the "combined" section of the WWFD tables (right-most columns of Figures 1 and 2) into a linear predictor to specify the corresponding linear model for each scenario (Stroup, 2013). For completeness and also to facilitate practical implementation, programming pseudo-code for the GLIMMIX procedure of SAS software (SAS Institute Inc., Cary, NC) specifying a general linear mixed model for scenarios A and B are included as Appendix A and Appendix B, respectively. Proper specification of a statistical model for data analysis is a rigorous process for which it is critical to have a strong grasp of hierarchical data structure by way of an in-depth understanding of the data collection process. Note that both experiments outlined in scenarios A and B have the general design outline of a 3×3 Latin square. Yet, differences in their randomization process lead to striking discrepancies in the hierarchical data structure relative to the specific treatment of interest, thereby identifying different physical entities as the actual experimental units (i.e., pen in a given period for scenario A and individual cow within a pen for scenario B).

In the interest of fulfilling the objectives stated for this communication, we now align our proposed scenarios A and B with the models proposed by Robinson (2016). Robinson's model 1 is appropriate given scenario B, whereas Robinson's model 2 follows from scenario A (Robinson, 2016). As noted before, scenarios A and B are not interchangeable and neither are their corresponding models; therefore, one cannot discuss either model without also giving context about the experimental design and the data collection process. Robinson fails to give a clear description of how the study was conducted, thereby making the decision between models 1 and 2 impossible. A clear and detailed description of how the data were collected and how the design was implemented is imperative for model specification. The inferential implications of disregarding experimental design and specifying a statistical model that does not match the process of data collection are not minor. If scenario A were to be improperly modeled with Robinson's model 1 (i.e., incorrectly treating cow as the experimental unit), one can anticipate at least 2 consequences of inferential relevance: (1) the denominator degrees of freedom for the *F*-test statistic on treatment would be artificially enlarged from 2 to 742 (Figures 1 and 2); and (2) the denominator of the corresponding *F*-test statistic would likely be somewhat decreased, thus inflating the corresponding *F*-ratio. These would, in turn, inflate type I error and thus increase the chances of false positives (Milliken and Johnson, 2009; Stroup, 2013). In other words, improperly specifying model 1 for scenario A would, on average, lead the researcher to conclude on more treatment differences as being statistically significant than should be.

Undoubtedly, subtle changes in how an experiment is run can have profound effects on how the statistical model for data analysis is specified (Milliken and Johnson, 2009; Stroup, 2013). The bottom line is that the statistical model should describe a plausible process that gives rise to the observations by (1) capturing the important independent variables affecting the outcome variable, and by (2) specifying any restrictions in randomization or any other data structure inducing correlation among observations. Mixed models are a statistical framework uniquely suited to this job (Littell et al., 2006; Milliken and Johnson, 2009; Stroup, 2013), provided that they are properly implemented. The inherently hierarchical structure of mixed models can naturally accommodate data with a hierarchical structure; that is, mixed models can properly "see" animal-level data, even in cases in which the animal is not the experimental unit. Further, in the context of mixed models, there is no need to collapse animal-level data into pen-level summaries, neither to "drop" animal-level data nor to "destroy" cow-level variance, as suggested by Robinson (2016). As such, mixed models can simultaneously recognize multiple sources of random variability in a data set, thereby assessing cow-level variability and pen-level variability at the same time, and using one or the other as experimental error for a given treatment of interest, as appropriate. Thus, mixed models can be used to ensure that the experimental unit for each treatment of interest is properly recognized within a study design, leading to proper recognition of the level of experimental error and thus to appropriate hypothesis testing. It is the estimated variation between the independent experimental units that determines the proper experimental error to assess treatment effects. Indeed, it is the estimated variance among experimental units that determines the estimated standard error of treatment differences (**SED**) that is used for classical hypothesis testing and ultimately, for elicitation of *P*-values. As worthwhile clarification, recall that it is the estimated SED, not the estimated standard error of the mean (**SEM**), that plays the meaningful role when testing for differential treatment effects. In fact, if there is any sort of hierarchical structure to the data (i.e., studies more complicated than a completely randomized design), the SEM is of no use for hypothesis testing (Littell et al., 2006; Milliken and Johnson, 2009; Tempelman, 2009; Stroup, 2013). This fact makes irrelevant points of Robinson's comparison of SEM between model 1 and model 2 (Robinson, 2016).

Notably, if cow were inappropriately treated as the experimental unit in a pen-based study as in scenario A, the SED for comparison of dietary treatments would be badly understated because it would ignore pen-to-pen variability (Tempelman, 2009).

Additional implications of inappropriate identification of the experimental unit of a study and improper specification of mixed models are broad in reach and encompass the important issue of scope of inference. In other words, what is the population to which any conclusions derived from a given study is meant to be applicable? Ideally, any experimental units used in a study would be considered a representative, if not a random, sample of a conceptual population of such units to which the conclusions are intended to apply. Hierarchical issues (i.e., cow, pen, herd) similar to those described for data structure apply here as well, such that the scope of inference of a study may be local (within a herd), regional (across herds in a given area), national or international, depending on the structure of data sources and the corresponding specification of random effects in mixed models for data analysis. An explicit discussion of scope of inference is relevant here for 3 reasons. First, a larger scope of inference, and thus a broader applicability of conclusions, is often accommodated by adding layers to the hierarchical design (e.g., on-farm studies repeated at several farms or experimental stations). Second, a sound experimental design should balance the allocation of experimental units and observational units such that meaningful results can be obtained at both smaller and larger scales of inference (i.e., cow level, pen level, herd level). For instance, if the variability between farms in their response to given treatments is considerable, it is important to characterize how farms differ, in addition to what occurs within individual farms. Last, scope of inference is likely to have implications on the seemingly pervasive issue of "research irreproducibility" across scientific disciplines that has been raised in multiple editorials published over the last several years (Ioannidis, 2005; Begley and Ellis, 2012; Nuzzo, 2015; Open Science Collaboration, 2015). Arguably, numerous problems are recognized as contributors to research irreproducibility, including unaccounted biases, "fishing expeditions" without adjustments for multiple testing, disregarded modeling assumptions, hypothesis myopia, asymmetric attention to unexpected results, and "just-so" storytelling, among others (Ioannidis, 2005; Nuzzo, 2015). Given the discussion presented thus far on the importance of properly characterizing data structure when specifying models for data analysis, one wonders whether it is possible that a limited appreciation, or even misunderstanding, of scope of inference might be misleading scientists to overgeneralize research results from a study that supports only a far narrower scope of inference? If so, failure to reproduce results should not be surprising. Specifically in the context of the dairy sciences, we have shown how a disregard for hierarchical data structure can inflate degrees of freedom and *F*-ratios, thereby leading to an unduly high rate of false positives that, not surprisingly, fail to replicate. A conscious effort to recognize proper scope of inference is needed to provide context within which to interpret results of any given study so that the general and specific circumstances for reproducibility of research results can be delineated. Statistically significant results may indeed be valid but only for limited scopes of inference (e.g., within a farm or region but not necessarily applicable more broadly) or under specific constrained conditions (e.g., if there were uncharacterized interactions of treatment with other factors). That is, research results on the effect of a given treatment, even if well characterized and tested in one setting, may not be applicable in a different context. A continued discussion on the reproducibility of research findings (or lack thereof) is relevant as scientists keep refuting themselves, resulting in confusion and disappointment among the general public, which in turn is likely to undermine public trust and discourage funding allocation for future research (American Association for the Advancement of Science, 2016). A closer consideration and better understanding of scope of inference might help explain, at least in part, research findings that are not reproducible. Further discussion on the issue of scope of inference in the context of the dairy sciences is provided by Tempelman (2009).

In closing, it is interesting to note that the same physical entity, either animal or pen, can play the role of an experimental unit or that of an observational unit within the same experiment (e.g., split-plot designs). In most cases, the existing general principles of design of experiments are such that virtually every quantitatively trained dairy scientist should be able to apply them to recognize the proper experimental unit for a treatment in a given pen-based animal study, be it cow, pen, or something else altogether. Proper case-specific application of these principles is important because the question of "what is the experimental unit" cannot always be reduced to a same answer for all treatment factors under all conditions. Proper identification of the experimental unit in a given study needs to be framed in the context of a specific research question—what is the effect of a treatment on a specific outcome of interest?—in combination with the many logistical nuances of the data collection process (i.e., experimental design and corresponding data structure). The same research question could very well have animal as the experimental unit in a tie-stall study (that is, assuming proper randomization of treatments to cows, even if cows are

fed separately), and pen as the experimental unit in another study addressing the same question but now in the context of animals confined to pens. Given the many logistical scenarios in which research questions in the dairy sciences can be posed, it is difficult, if not impossible, to outline "be-all and end-all" guidelines defining what is the experimental unit for every possible pen-based animal study; instead, statistical expertise should be sought and applied on a case-by-case basis. The need for such specialized statistical expertise raises awareness of the importance of modern quantitative training for the next generation of dairy scientists (i.e., our current graduate students), as well as continuing quantitative education of established dairy researchers, journal editors, and reviewers. Most importantly, it is worth emphasizing that specification of the experimental unit in a given experiment is not a matter of opinion; rather, it is determined by how the experiment was set up, how the data were collected, and the intended scope of inference. Untangling logistical nuances and interpreting their implications in an experimental setting may require tailored expertise in experimental design. In our opinion, engaged collaborative interdisciplinary research interactions between dairy scientists and statisticians hold the key to ensuring efficient and powerful science that is both reproducible and replicable in the real world and thus ultimately relevant to stakeholders and to the public.

REFERENCES

American Association for the Advancement of Science (AAAS). 2016. Historical Trends in Federal R&D. AAAS, Washington, DC. Accessed May 12, 2016. http://www.aaas.org/page/historical-trends-federal-rd#Agency.

Begley, C. G., and L. M. Ellis. 2012. Raise standards for preclinical cancer research. Nature 483:531–533.

Casella, G. 2008. Statistical Design. 1st ed. Springer Texts in Statistics. Springer, Gainesville, FL.

Ioannidis, J. P. 2005. Why most published research findings are false. PLoS Med. 2:e124.

Kuehl, R. O. 2000. Design of Experiments: Statistical Principles of Research Design and Analysis. 2nd ed. Brooks/Cole, Cengage Learning, Belmont, CA.

Lamberson, W. R. 2016. Letter to the editor: A response to Robinson (2016). J. Dairy Sci. 99:2437.http://dx.doi.org/ 10.3168/jds.2016-10915.

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.

Milliken, G. A., and D. E. Johnson. 2009. Analysis of Messy Data. Vol. 1: Designed Experiments. 2nd ed. Chapman & Hall/CRC Press, Boca Raton, FL.

Nuzzo, R. 2015. Fooling ourselves. Nature 526:182–185.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science 349(6251):4716. 10.1126/science.aac4716.

Robinson, P. H. 2016. Letter to the editor: Comments on *Journal of Dairy Science* statistical unit policy as it related to pen-based animal studies. J. Dairy Sci. 99:2435–2436. http://dx.doi.org/10.3168/jds.2015-10712.

Stroup, W. W. 2013. Generalized Linear Mixed Models: Modern Concepts, Methods and Applications. 1st ed. Texts in Statistical Science. Chapman & Hall/CRC Press, Boca Raton, FL.

Stroup, W. W. 2015. Rethinking the analysis of non-normal data in plant and soil science. Agron. J. 107:811–827.

Taylor, C. C., and M. S. Allen. 2005. Corn grain endosperm type and brown midrib 3 corn silage: site of digestion and ruminal digestion kinetics in lactating cows. J. Dairy Sci. 88:1413–1424.

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(E-Suppl.):E162–E172.

Tempelman, R. J. 2009. Invited review: Assessing experimental designs for research conducted on commercial dairies. J. Dairy Sci. 92:1–15.

Appendix A

The SAS code below corresponds to specification of a general linear mixed model for scenario A, Treatments independently assigned to pens within a period, following from the "What Would Fisher Do?" (WWFD) exercise in Figure 1. This SAS code assumes a (conditional) normal distribution on the response variable y and an experimental design consisting of a single (i.e., unreplicated) Latin square:

```
proc glimmix data=dataset plots=studentpanel;
       class Pen Period Trt Cow;
       model y = Trt;       * The response variable y can be entered as observed, that
                            *   is, at the level of individual cows measured at each
                            *   period without any need to "collapse" or "drop"
                            *   cow-level data;

       random Pen Period;   * Random effect terms for pen and period, as standard for a
                            *   single 3x3 Latin Square design;
       random Pen*Period*Trt;     * Random effect term to identify the pen in a given
                                  *   period as the experimental unit for Trt. This
                                  *   corresponds to the term pen*period|treatment in
                                  *   the "combined" column of the corresponding WWFD
                                  *   exercise (Figure 1);

       random Cow(Pen);     * Random effect term to identify cow as a unit of
                            *    subsampling within a pen;

       * The residual term is specified by default at the level of observation,
       *       in this case corresponding to the term period x cow(pen) in the
       *       corresponding WWFD exercise (Figure 1);

       * Note: the order of presentation of the random statements is inconsequential
       *       to model specification.
run;
```

If period were to be treated as an element of the treatment structure in the WWFD exercise, as is sometimes legitimate for dairy applications (Tempelman, 2004), the corresponding SAS code should be modified as follows to reflect the appropriate mixed model specification:

```
proc glimmix data=dataset plots=studentpanel;
       class Pen Period Trt Cow;
       model y = Trt Period;
       random Pen;
       random Pen*Period*Trt;
       random Cow(Pen);
run;
```

Further, for replicated Latin square designs, researchers should also incorporate into the model the interaction between treatment and period (Tempelman, 2004), either as a random effect or a fixed effect, consistent with the specification of period as an element of design or treatment structure (Stroup, 2013), respectively.

Appendix B

Accompanying SAS code corresponding to specification of a general linear mixed model for scenario B: Treatments independently assigned to cows within a pen-by-period combination, following from the WWFD exercise in Figure 2. This SAS code assumes a (conditional) normal distribution on the response variable y and a single (i.e., unreplicated) Latin square design:

```
proc glimmix data=dataset plots=studentpanel;
       class Pen Period Trt Cow;

       model y = Trt;        * The response variable y can be entered as observed, that
                             *   is, at the level of individual cows measured at each
                             *   period without any need to "collapse" or "drop"
                             *   cow-level data;

       random Pen Period;    * Random effect term for pen and period, as standard for a
                             *   single 3x3 Latin Square design;

       random Pen*Period;    * In this case, the Pen*Period term identifies the Pen-at-
                             *   each-Period, which effectively acts as a blocking
                             *   factor within which treatments are randomly allocated
                             *   to cows;

       random Cow(Pen);      * Random effect term to identify cows within pens as a
                             *   source of random variability;

       * The residual term is specified by default at the level of observation,
       *       in this case corresponding to the term "period x cow(pen) | treatment"
       *       in the corresponding WWFD exercise (Figure 2) and identifies the
       *       experimental unit for Trt;

       * Note: the order of presentation of the random statements is inconsequential
       *       to model specification.
run;
```

If period were to be treated as an element of the treatment structure in the WWFD exercise, as is sometimes legitimate for dairy applications (Tempelman, 2004), the corresponding SAS code should be modified as follows to reflect the appropriate mixed model specification:

```
proc glimmix data=dataset plots=studentpanel;
       class Pen Period Trt Cow;
       model y = Trt Period;
       random Pen;
       random Pen*Period;
       random Cow(Pen);
run;
```

Further, for replicated Latin square designs, researchers should incorporate into the model the interaction between treatment and period (Tempelman, 2004), either as a random effect or as a fixed effect, consistent with the specification of period as an element of design or treatment structure, respectively (Stroup, 2013).