

Improved Experimental Design and Analysis for Long-Term Experiments

Thomas M. Loughin*

ABSTRACT

This paper addresses inadequacies in the way most long-term experiments (LTEs) are conducted and analyzed. The standard design under which LTEs are usually conducted involves a *fixed start*, establishing all plots in the study in the same year. This design is shown to be inadequate for the purpose of testing and estimating the time \times treatment (TRT) interaction, which is generally the primary interest in a LTE. This inadequacy occurs because the repeated measures taken on every plot are all influenced simultaneously by the same random environmental conditions, the effects of which are confounded with the fixed effects of interest. No statistical analysis can completely separate the fixed effects from the random nuisance effects, although added assumptions about the shape of trends across time or covariates to describe the random effects can sometimes be helpful. An alternative experimental design, the *staggered-start* design, has been used to alleviate this confounding by establishing plots from different blocks in successive years, but proper analysis of this design has not been presented. A correct analysis of the staggered-start design is determined and presented. The analysis is applied to hypothetical data from a staggered-start design whose true means are known, and it is shown to do a much better job of estimating these means than any methods applied to data from the standard design. A staggered start should be considered instead of a fixed start for all future LTEs.

LONG-TERM EXPERIMENTS are often conducted to compare long-term effects (e.g., sustainability) of various TRTs on one or more response measurements. Examples include comparisons of different amounts of fertilizers on soil fertility and other properties (Aulakh et al., 1991; Barber, 1979; McCollum, 1991), comparisons of tillage effects on crop yields (Bailey et al., 1996), studies of pest control (Alldredge and Young, 1995; Boström and Fogelfors, 2002), and, of course, the classics at Rothamsted (Johnston, 1994; Poulton, 1996a, 1996b). The common defining trait of all LTEs is the assignment of TRT regimes to all plots at the start of the experiment and subsequent repeated measurement of the treated plots over a span of many years. Durations of LTEs naturally depend on the goals of the research, and may range from just a few years (sometimes called *feasibility studies*; see McRae and Ryan, 1996) to well over 150 years are still ongoing.

Long-term experiments can be arranged according to any valid experimental design, including randomized block and split-plot designs. It is evidently the standard that all plots are initiated at a single starting time, although in some crop rotation experiments, plots may be

phased in as needed to allow each crop of the rotation to be grown in each year of the study (Cochran, 1939; Yates, 1954; Patterson, 1964). Measurements taken on the plots are generally taken each year in the case of crop yield and other plant measurements, although some measurements can be taken more frequently (soil characteristics, for example) and some less frequently (yields of a specific crop in a rotation, for example). These *repeated measures* data are ultimately subjected to some kind of statistical analysis, often with the goal of understanding something about the potential for different cumulative effects of TRTs over time. As a result, the time \times TRT interaction often becomes the focus of an analysis. Alternatively, analyses at specific, selected time endpoints are sometimes conducted.

It is generally accepted that measurements taken on field experiments may be influenced by uncontrollable environmental factors. Furthermore, these factors may impact some TRTs in a study differently from others. For example, drought-resistant crops tend to yield better under limited rainfall conditions than susceptible varieties, but the yield relationship may be reversed in years with ample moisture. It is therefore quite likely that, in addition to any fixed, repeatable time and time \times TRT effects present in the measured responses, there are also uncontrollable, random year-to-year fluctuations and year \times TRT random effects. (In this paper, *time* is used to mean the period elapsed since the commencement of TRT application in an experimental plot, and *year* is used to mean the calendar year in which measurements are taken.) In many field trials, measuring these random effects is not a particular research interest. Rather, they are often considered a nuisance. Because they are natural and cannot be prevented, careful experimental design and analysis must be done in order to prevent these nuisance effects from interfering with the real research goals of understanding the fixed effects.

Commonly-used methods of designing and analyzing LTEs in the plant sciences do not adequately account for the random effects associated with years or any year \times TRT interaction. In fact, it is shown in Loughin et al. (2006) and reiterated in this paper that the standard method of designing LTEs in field trials—initiating TRT application on all plots at the same time—completely confounds (confuses) the fixed effects that represent the research goals with the random effects that are nuisance. Variations in the measurements or in the TRT differences across years are simultaneously due to both the fixed and the random effects, and the two sets are inseparable from each other. No valid statistical analysis of such experiments can be performed that can

Dep. of Statistics and Actuarial Science, Simon Fraser Univ., Burnaby, BC, Canada V5A 1S6. This work was done while the author was on faculty in the Department of Statistics at Kansas State University. Received 27 Apr. 2006. *Corresponding author (tloughin@sfu.ca).

Published in Crop Sci. 46:2492–2502 (2006).
Analysis of Unreplicated Experiments (Symposium)
doi:10.2135/cropsci2006.04.0271

© Crop Science Society of America
677 S. Segoe Rd., Madison, WI 53711 USA

Abbreviations: AIC, Akaike Information Criterion; LTE, long-term experiment; TRT, treatment.

automatically isolate and analyze the fixed effects and lead to valid interpretations about them. Only through additional assumptions can analyses be constructed that have the potential to separate the fixed and random effects, but each assumption must be completely true in order for the separation to be complete. Unfortunately, none of the assumptions can be verified through tests or investigations of the data, so each must be made on faith and can readily be called into question. Loughin et al. (2006) discusses these issues at some length from a more statistical perspective. In the present paper, these assumptions are discussed more practically as they apply to agronomic research, and implementation of the resulting analyses is described in greater detail.

The inadequacy in this standard LTE design is demonstrated in this paper to be its lack of repetition of the environmental factors to which the treated plots are exposed. Achieving true replication in time in the usual way—repeating the entire experiment multiple times, each in succession upon conclusion of the previous replicate—is not feasible with LTEs, whose duration may be many years. A good compromise that achieves replication with little time cost is to stagger the start of TRTs in different blocks across years. That is, blocks within the experiment are run by starting Block 1 in the first year, Block 2 in the second year, and so forth for as many replicates as are planned in the research. This staggered-start design has been mentioned many times before, both in crop sciences (Smith, 1979; Preece, 1986; McRae and Ryan, 1996; Martin et al., 1998; Orchard et al., 2000) and in other disciplines where LTEs are done (Walters et al., 1988). Judging from a survey of the crop science literature, however, this design does not appear to have been adopted by many, if any, researchers. It may be speculated that this is because the substantial drawbacks of the standard LTE design and corresponding benefits of the staggered-start design are not well-understood by the researchers who design, conduct, and analyze LTEs. It is hoped that this paper, along with a companion paper in the Statistics literature (Loughin et al., 2006) can correct this oversight, and that scientists initiating future LTEs will do so using a design that permits proper analysis and interpretation of the results.

PROBLEMS WITH THE STANDARD DESIGN OF LTEs

We consider the basic generic problem of comparing the long-term effects of t TRTs in a randomized complete block (RCB) design with n complete blocks for a period of r observation times, and we suppose that these observations are taken annually, such as a crop yield. (All of the problems and solutions discussed in this paper carry over completely to experiments with alternative experimental designs such as split-plots, alternative TRT structures such as factorials, and alternative measurement periods.) Most LTEs are designed in such a way that all treated plots are initiated at the same time. This results in a design which is depicted in Fig. 1. Treatments are randomized to plots in each block at the

Block	TRT	Year 1	Year 2	Year 3	Year 4	Year 5	etc.
		Time 1	Time 2	Time 3	Time 4	Time 5	etc.
1	2						...
	5						...
	1						...
	4						...
2	3						...
	1						...
	4						...
	3						...
3	5						...
	2						...
	5						...
	4						...

Fig. 1. Schematic of the standard design of a long-term experiment.

start of Year 1, and at the end of this year the measurements for the first time (Time 1) are taken on all plots. Time 2 measurements are all taken at the end of Year 2, Time 3 at the end of Year 3, and so forth for the duration of the study.

From this description, it is apparent that any environmental factors affecting the measurements made in a given year affect all plots simultaneously. In a favorable season, all plots will yield well, and conversely. Furthermore, and perhaps more importantly, TRTs that respond unusually well or unusually poorly to the conditions present in a given year will do so in all blocks simultaneously. The problem with this is that any analysis conducted on the measurements from that year is not able to distinguish the repeatable, fixed effects of the TRTs at that particular time from the uncontrolled and unrepeatable random effects from that one year. For example, the direction of a comparison between drought-resistant and drought-susceptible varieties of a crop would change depending on whether it is a particularly wet or dry year, and the magnitude of this random effect may overwhelm that of any long-term mean trends that may be present.

This is obviously not a good thing. The goal of most experiments is to make statements about the general (future) potential of the TRTs rather than to understand how they reacted once in an unrepeatable past. In order to do this, TRT comparisons need to be made across a sampling of the environments to which the conclusions are to be applied. This is well known, and in fact is the impetus behind the advice given in the Instructions to Authors for *Agronomy Journal* (www.asa-cssa-sssa.org/publications/pdf/ajinstrauthor.pdf; verified 29 Sept. 2006): “Field experiments that are sensitive to environmental interactions and in which the crop environment is not rigidly controlled or monitored, such as studies on crop yield and yield components, usually should be repeated (over time or space, or both) to demonstrate that similar results can, or cannot, be obtained in another environmental regime.” If we were to ignore the “long-term” aspect of a LTE, however, and just analyze the results from Time 1, we would have data from only

the environment observed in Year 1, which would be contrary to this advice.

Extending this notion, in any LTE designed according to Fig. 1, the entire set of repeated measures made on each plot is exposed to the same sequence of environmental factors. Unusual or adverse events in the first year of the study could have random, residual effects well beyond the first year, for example. These events might not occur were the LTE established in a different year. There is no way to demonstrate that whatever apparent long-term TRT effects were observed could have been obtained under any other sequence of environmental effects. Interpretation of the analysis results cannot be extended to make inferences to any other sequence of years other than the one under which the experiment was conducted.

To make ideas clearer, consider the hypothetical example represented by Fig. 2a and 2b. Figure 2a represents the true means of four different TRTs from a LTE conducted for 5 yr. These are the numbers that would be observed on all plots of the experiment if variability did not exist. Now consider what happens when random year and year \times TRT variability is added to the means. Figure 2b depicts one possible outcome based on a hypothetical sequence of random effects. Notice that all means rise above and fall below their true

values simultaneously due to high annual variation. This phenomenon is not unusual: Bailey et al. (1996) describe an experiment in which annual variations accounted for 92% of the total variation in corn yields in a LTE. Martin et al. (1998, p. 30) report, "...the major source of variation in crop performance in Australia is the climate or year effect."

Now suppose that we wanted to estimate or predict the mean response after two years of TRT using data exhibiting this kind of variability. An analysis of Time 2 data here would result in a substantial underestimate of the true mean at Time 2. This highlights the problem that is laid out more mathematically in Loughin et al. (2006): no analysis of data from a LTE following the design represented in Fig. 1 can be relied upon to provide accurate answers to common research questions, unless further assumptions are made about either the patterns of the true means across time or the nature of the random effects.

IMPROVED STATISTICAL ANALYSIS OF LTEs USING THE STANDARD DESIGN

Standard Analysis of Repeated Measures Data

Data from a LTE are in the form of repeated measurements on each treated plot. Even without the unique

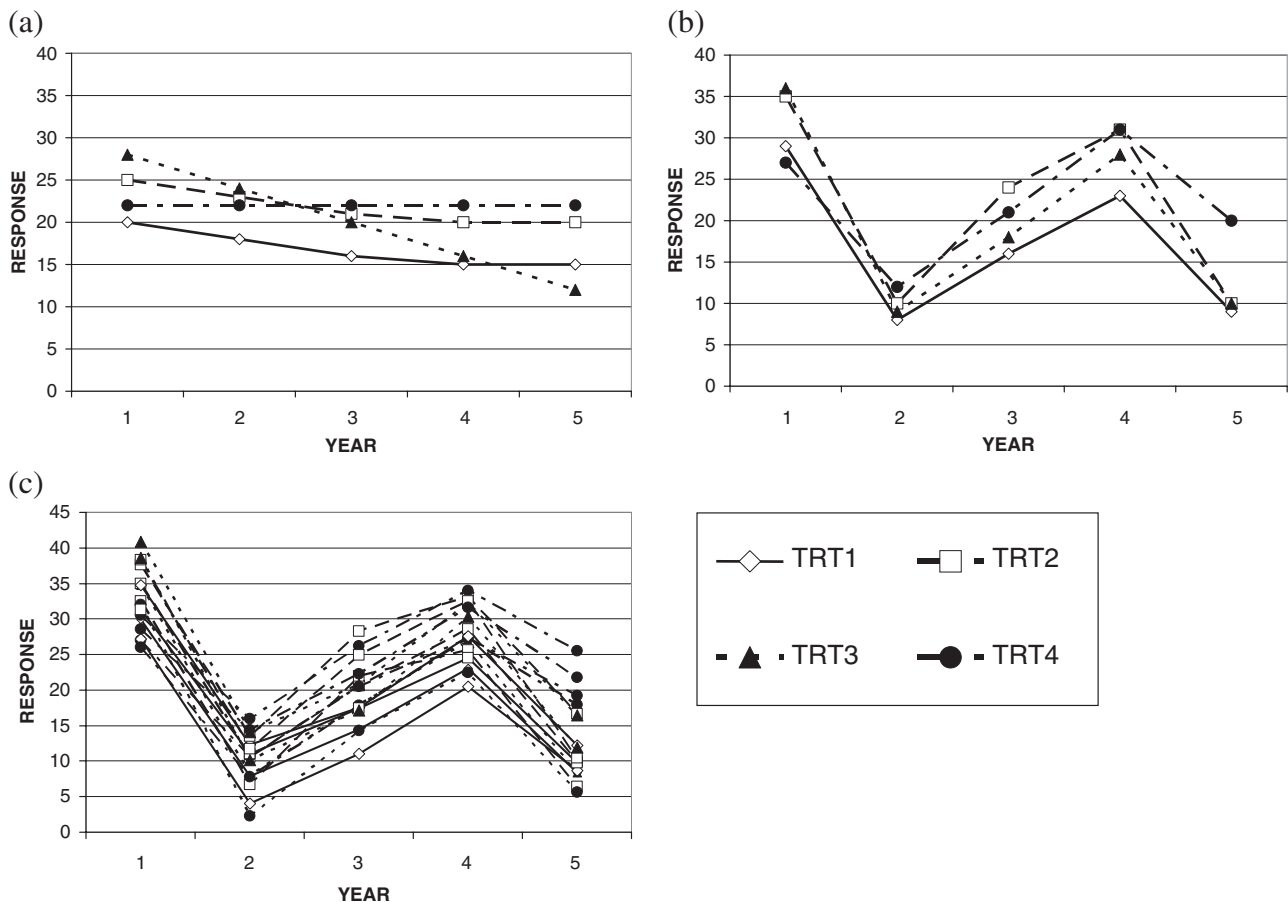


Fig. 2. Artificial example of data generated according to a hypothetical long-term experiment using the standard design with four treatments in four blocks measured for 5 yr. (a) True treatment means at each time corresponding to fixed effects only. (b) Treatment means at each time following addition of random effects for year and year \times TRT. (c) Four blocks of data generated from the means in (b).

problems associated with LTEs, repeated-measures data present a special challenge for statistical analysis. Webster and Payne (2002) provide a very nice overview of valid statistical analysis methods for repeated-measures data. Among the most common approaches are (i) the *summary statistics* approach (also called *response feature* or *derived variables*, see Mead, 1988), (ii) multivariate ANOVAs, and (iii) modeling the correlation structure. A brief summary of these procedures follows.

Summary Statistics

The summary statistics approach to repeated measures is to reduce the multiple measurements on each plot to one or more summary statistics that measure some phenomenon of interest. For example, suppose that the response measurement is crop yield. Then the mean (or total) yield taken across all times could be used to compare the overall productivity of plots with different TRTs. The maximum or minimum yield or the range in yield on each plot could be used to compare of stability of yields across time among TRTs. A slope of the yield measurements across time could be used to determine whether TRTs follow similar trends. Differences between consecutive years, the difference between the last and first year, and measurements taken at individual years are additional examples of summary statistics. The point to make about this form of analysis is that, regardless of the summary being used, the calculation is performed separately and identically on each plot, reducing the r years of measurements to a single summary statistic. These summary statistics are then analyzed in accordance with the design of the experiment as if it were the only measurement made on the each plot. For example, if the TRTs are assigned in a RCB design, then the summary statistics are analyzed using a RCB ANOVA.

This analysis approach has the advantage of being relatively easy to construct and interpret. One can analyze as many different derived variables as are needed to compare the TRTs with respect to all research questions. Each analysis provides a straightforward answer to the question implied by the derived variable. There are relatively few statistical assumptions (primarily only those normally used in conjunction with ANOVA), and there are well-known methods for checking those assumptions. There are some drawbacks to this approach. First, one gets answers only to those questions that one poses through the summary statistics. If, for example, there is a substantial difference in yield ranges among the TRTs, but the range was not one of summary statistics considered, then this difference may never be discovered. Another potential drawback is the inflation of the type I error (rejecting true hypothesis) rate that comes with performing many tests at once. If formal control of the type I error rate is required, then this can be achieved easily using the method of Bonferroni (covered virtually anywhere multiple testing is discussed; see Ott and Longnecker, 2001). Using the Bonferroni method, one does not declare a test to be

significant unless its $P < (\alpha/g)$ where g is the number of summary statistics that were created. This method can be excessively conservative, however—it often fails to find TRT effects that are real—and so it is recommended for use only if an experiment can be designed to have sufficient power to overcome this (i.e., ample replication). Alternatively, we can attempt to address potential inflation of the type I error rate informally by cautious interpretation of marginal significance seen in a small number of tests out of many done.

Multivariate Analysis

Repeated measurements across time that are taken on a single plot can be viewed as a single, multivariate response measured on that plot. A multivariate analysis simultaneously models the r response measurements according to the design of the experiment. Effects of TRTs are compared based on best combinations of response variables as determined by correlation structure among them. A single test simultaneously compares all TRTs at all times.

While this universal test is convenient, this procedure often lacks power to detect real differences. This difficulty is exacerbated when the number of times is large relative to the degrees of freedom for error in the original design, and the method cannot be used at all when the number of times is greater than the number of error degrees of freedom. This method is rarely used in the analysis of repeated measures.

Modeling the Correlation Structure

One analysis approach that has gained popularity in recent years is to model the serial correlation associated with the repeated measures, and then to base inferences (tests, contrasts, estimates of means, standard errors, and confidence intervals) on a mixed model ANOVA that incorporates the estimated serial correlation structure. *Serial correlation* here refers to the correlation between measurements taken on the same plot across time.

Cochran (1939) recognized that LTE data looks like data from a split-plot design, if one considers repeated measurements of a subject at different times as being similar to repeated measurements of a whole-plot unit at different levels of a subplot factor. He also recognized, however, that serial correlations between measurements taken relatively close together in time are likely to be greater than those taken far apart in time. He therefore did not actually recommend analysis of repeated measures as if they were from a split-plot design, because the latter implicitly assumes that correlations between all pairs of times are equal. Nonetheless, use of split-plot analysis for repeated-measures data has flourished, largely because there were no commonly-available computational tools with which to incorporate varying levels of serial correlation between different times.

More recently, computing capabilities have improved to the point that modeling the serial correlation among repeated-measures data is now possible for anyone with access to certain standard software packages. PROC MIXED in SAS, for example, has extensive capabilities

for modeling serial correlations according to numerous potentially-feasible structures. Guerin and Stroup (2000) recommend selecting a structure by fitting a large number of these structures to the data and choosing the one with the smallest Akaike Information Criterion (AIC) statistic. The authors of *Analysis of Messy Data Volume 1: Designed Experiments* (Milliken and Johnson, 1992) are presently revising this book and indicate that they, too, will recommend using AIC to select a correlation structure for repeated measures analysis (G.A. Milliken, 2006, personal communication).

This approach to analysis has several advantages. It provides formal tests of time, TRT, and time \times TRT, which the other methods do not do; all time \times TRT combinations can be compared in any way through selected contrasts; and there is a potential gain in power for TRT comparisons when the correlation model is chosen correctly. There are also some disadvantages to this approach, not the least of which is computational complexity. Especially when data are measured on many times, there is the potential for some models to give a poor fit or no fit at all. When this happens, then either some advanced programming skills are needed to coerce a better fit of the affected models, or the selection of a best structure must be made from only those that are easiest to fit. There are also more statistical assumptions involved in this procedure; in particular, one must assume that the structure chosen by the data is, in fact, the true structure for the population from which the data were sampled. There is no way to check this assumption, although the simulation results of Guerin and Stroup (2000) suggest that failure of this assumption does not generally have a severe adverse effect on the quality of the analysis as long as the selected structure is a reasonable approximation to the correct one. They also indicate that the use of AIC as a selection criterion seems generally to provide adequate approximations.

Hypothetical Example

Because this method serves as the basis for later analyses, we examine an example in some detail. Consider the hypothetical example from Fig. 2a and 2b. (Using hypothetical data allows us to compare observed analysis results with the known truth, something we cannot do with a typical real example.) Figure 2a represents the true means that are the target of a LTE. Figure 2b shows those same means, altered by some random year and year \times TRT effects. From these means, data were generated in four blocks by adding random block, block \times year, and Residual effects to four copies of the means. Sizes of random effects were chosen to mimic the magnitudes of effects observed from the analysis of real LTE data from the study described in Schlegel and Havlin (1995). The final data are shown in Fig. 2c.

A repeated-measures analysis incorporating models for the correlation structure is conducted starting with the ANOVA shown in Table 1. The structure is that of a split-plot design with TRT as the whole plot factor and time as the subplot factor. The SAS code also given in

Table 1. Repeated-measures ANOVA and SAS code for a randomized complete block design analysis incorporating models for the correlation structure. Variable names in the program are in capital letters.

Source†	df	SAS code for analysis‡
Block	$n - 1$	<code>proc mixed method=reml;</code> <code>class BLOCK TRT TIME;</code> <code>model Y = TRT TIME TRT*TIME /</code> <code>ddfm=kr;</code> <code>random BLOCK;</code> <code>repeated TIME / subject=</code> <code>BLOCK*TRT type = _____;</code> <code>lsmeans TRT*TIME / diff;</code> <code>contrast ...;</code> <code>run;</code>
TRT	$t - 1$	
Block \times TRT	$(n - 1)(t - 1)$	
Time	$r - 1$	<code>repeated TIME / subject=</code> <code>BLOCK*TRT type = _____;</code>
Time \times TRT	$(r - 1)(t - 1)$	
Error	$t(r - 1)(n - 1)$	
Total	$ntr - 1$	

† TRT, treatment.

‡ Options for “type=” are described in the text. Lsmeans and contrast statements vary according to research needs.

Table 1 shows how this analysis can be conducted using PROC MIXED. The important feature that distinguishes the repeated measures analysis from that of a split-plot is the REPEATED statement, wherein different models for the correlation structure can be proposed. The “subject” identified in the REPEATED statement is the unit upon which the repeated measurements are taken. In this case, a plot is identified uniquely by a combination of block and TRT. Because block \times TRT is the subject, it does not appear in the RANDOM statement as it normally would for a split-plot analysis. Different models for the serial correlation are considered through the TYPE= option. Commonly-considered structures include compound symmetry (CS), first-order autoregressive (AR(1)), Toeplitz/banded (TOEP), heterogeneous-variance versions of these (CSH, ARH(1), and TOEPH, respectively), first-order antedependence (ANTE(1)), and unstructured (UN or UNR, depending on whether you wish to see results expressed as covariances or correlations, respectively). See Littell et al. (1996), Guerin and Stroup (2000), and the built-in documentation for SAS (SAS Institute, 2004) for further details on the use of this statement and the different correlation models that can be fit.

Following the analysis recommendation in Guerin and Stroup (2000), we fit all of these correlation models to the example data and calculated their AIC values. Because the data were generated with no added correlation structure, compound symmetry is the correct structure for this example, and it does indeed turn out to produce the smallest AIC value (data not reported). Results of tests for fixed effects show that TRT, time, and time \times TRT are all highly significant (all $P < 0.0001$). Not surprisingly, the estimated means (Fig. 3) are quite close to those depicted in Fig. 2b. The patterns associated with the target means in Fig. 2a are not evident from the means estimated by this analysis, which confounds the fixed and random effects.

Assumptions to Improve Analysis of LTEs

One of the fundamental assumptions of any statistical analysis of repeated-measures data is that the subjects upon which the repeated measurements are taken—the plots in a LTE—must respond independently from one another. That is, knowing whether one plot’s measure-

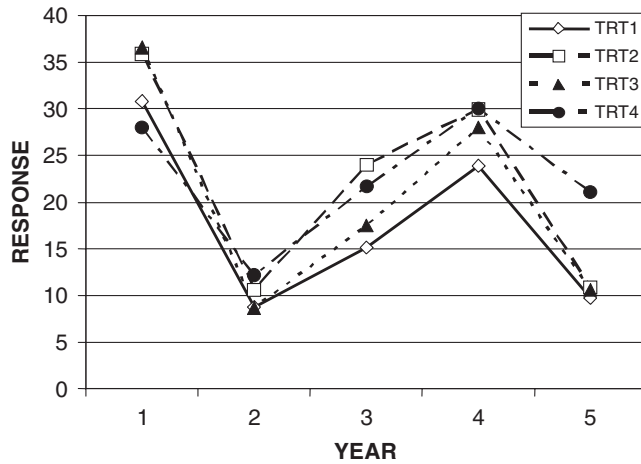


Fig. 3. Estimated treatment means at each time from the analysis of standard-design data without accounting for the random effects of year and year \times TRT.

ment is above or below the true mean should provide no information about where another plot's measurement might lie. This is not the case with LTEs employing the standard design, because all plots are affected simultaneously by the random effects of year. For example, if one plot is far above its true mean due to favorable conditions, then others from the same year are likely to respond high as well. Therefore, none of the analysis methods described in the Standard Analysis of Repeated Measures Data section are valid when applied to LTE data.

Ideally, this problem with lack of independence could be remedied by including random effects into the analysis to account for the random year and year \times TRT effects. Then the dependence described above would be explained by effects in the model, leaving a much less restrictive assumption that the plots respond independently after the year and year \times TRT effects are considered. This cannot be done with standard LTE data, however, because of the other problem noted in the Introduction that the random year and year \times TRT effects cannot be distinguished from the fixed time and time \times TRT effects. Including random effects related to years into the model would nullify the analysis of time and time \times TRT effects.

The only alternative for separating these fixed and random effects is to make some assumptions about their natures and to base a subsequent analysis on these assumptions. Loughin et al. (2006) offers three assumptions that can be used for this purpose to amend an analysis based on modeling the correlation structure. The two most practical of these are considered here in some detail. One makes an assumption about the structure of the fixed effects for time; the other makes an assumption to about the random effects for year. Both result in modifications to an analysis that models the serial correlation as described above.

Assumptions about the Fixed Effects

If it is anticipated that the means associated with the fixed time effects should follow a particular pattern, then

knowledge of that pattern can be incorporated into the analysis. The simplest, and perhaps most useful, trend to consider is the linear trend, although other models for trends, such as polynomial models or an exponential decay, might be more appropriate in particular problems. Assuming that the means for each TRT group follow *exactly* the prescribed pattern across time, then it can be concluded that any deviation of the observed TRT means around the time trend is due to randomness. Random-effect variance components can be estimated for year and year \times TRT as well as for any other random-effect terms as necessitated by the design of the experiment.

To see how this assumption works, we return to the hypothetical example. Suppose that, based on best-known principles, we believe that the means should approximately follow straight lines across time for each TRT. Notice from Fig. 2a that this is not quite correct, which represents the reality that we rarely know exactly what trend the means follow. Incorporating the straight lines into the ANOVA from Table 1 results in the ANOVA and SAS code in Table 2. Notice that because we are fitting a regression line across time for each TRT rather than separate, unrelated means at each time, the levels of time are now being considered as numerical rather than as classification variables. Also, time can no longer be in the REPEATED statement because it is no longer a classification variable. Instead, we arrange for the measurements to be sorted in time sequence for each plot and do not specify any variable as the repeated measures effect. SAS assumes that the ordering of the measurements in the data set represents the ordering of the times. Finally, we now can add random effects for year and year \times TRT whose variance components will be estimated separately from the assumed time trend.

This model was applied to the data above. Results of the AIC comparisons for different correlation structures indicate that compound symmetry is the best-fitting correlation structure. The tests for fixed effects are now

Table 2. Repeated-measures ANOVA and SAS code for analyzing long-term experiment data assuming linear trends for treatment means across time. Analysis assumes a randomized complete block design and incorporates models for the correlation structure. Variable names in the program are in capital letters.

Source [†]	df	SAS code for analysis [‡]
Block	$n - 1$	proc mixed method= reml data=set1; class BLOCK TRT YEAR; model Y = TRT TIME TRT* TIME / ddfm=kr; random BLOCK YEAR YEAR*TRT;
TRT	$t - 1$	
Block \times TRT	$(n - 1)(t - 1)$	
Time	1	repeated / subject=BLOCK*TRT type=_____;
Time \times TRT	$t - 1$	
Year	$r - 1$	lsmeans TRT / diff at TIME=1; lsmeans TRT / diff at TIME=2; lsmeans TRT / diff at TIME=3; lsmeans TRT / diff at TIME=4; lsmeans TRT / diff at TIME=5; contrast...
Year \times TRT	$(r - 1)(t - 1)$	
Error	$t(r - 1)(n - 1) - t$	
Total	$ntr - 1$	run;

[†] TRT, treatment.

[‡] Options for "type=" are described in the text. Lsmeans statements allow pairwise comparisons of means at each time.

much more subdued, reflecting the real fixed effects of time, which are much smaller than the combined fixed and random effects. The test for time \times TRT is significant at the 0.05 level ($F_{3,9} = 5.12$, $P = 0.02$), while those for TRT and time are not ($F_{3,9,09} = 3.22$, $P = 0.074$, and $F_{1,3} = 0.43$, $P = 0.56$, respectively). The estimated means from this analysis are shown in Fig. 4. Notice that these means much more closely resemble those from the previous analysis, indicating that this analysis approach has the potential to separate the fixed and random effects when the assumed trend is a reasonable fit for the true mean trend. On the other hand, comparing Fig. 2a and 4 reveals that certain more subtle features of the means, such as the leveling off of means for TRTs 2 and 3, are not captured by the linear trends analysis. This demonstrates the serious drawback of this analysis approach: when the assumed model is wrong, the answers may fail to identify important facets of the underlying truth, or they may make little sense at all if the model is drastically wrong. Unfortunately, the form of the assumed model must be chosen completely on faith, because there is no way to test whether it is appropriate.

Assuming polynomial trends would require only adding higher powers of time to the ANOVA table and to the MODEL statement. Other models would require more careful implementation.

Assumptions about the Random Effects

So-called random variations associated with environments can often be explained to some extent by measurements taken on those environments. Rainfall, for example, is correlated with many crop and soil characteristics. Other measurements, such as solar radiation or pest incidence, may also be related to random changes in the responses. Theoretically, if we had sufficiently detailed information, we might be able to explain 100% of the variation due to environments. Practically speaking, measurements that are typically available cannot achieve this goal, but may be able to come close. Therefore, another approach for improving

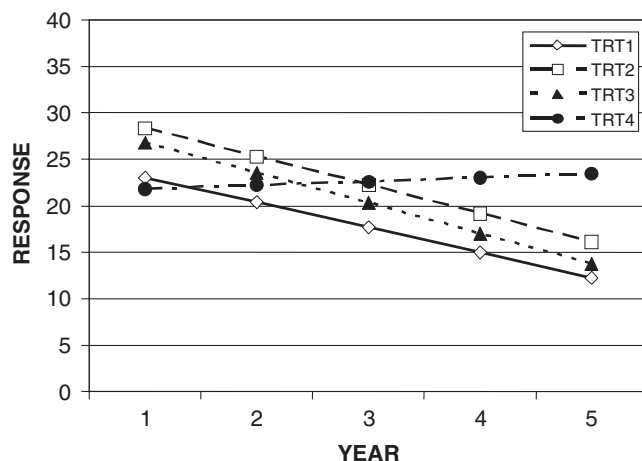


Fig. 4. Estimated treatment means at each time from the analysis of standard-design data using a linear trend model to represent the means of treatments across time.

the analysis of a LTE is to add covariates to the model to serve in lieu of the random effects of year and year \times TRT.

Specifically, suppose P different measurements, labeled x_1, x_2, \dots, x_P , are available as plot-year covariates, meaning that they are measured on each plot separately in each year (flood damage, pest load, and such). Note that it is important that these measurements cannot be anything that could conceivably have been influenced by the TRTs, which excludes things like annual soil N measurements in a rotation or fertilizer study, or pest load in a pest-management study. Plot-year covariates along with their interactions with TRT, can be included directly into the model as fixed effects. See Table 3 for an example using $P = 2$ plot-year covariates.

Conceptually, if the covariates happen to explain the random year and year \times TRT effects completely, then they separate the random effects from the fixed effects associated with time and time \times TRT, and an analysis conducted according to Table 3 provides legitimate inferences that are purely about these fixed effects. In practice, however, there are some difficulties in implementing this form of analysis.

First, the covariates must be measured separately on each plot in each year. Covariates that take the same value for all plots in a given year, such as measurements of rainfall from a single rain gauge, cannot be used in this form of analysis. This is because such covariates are partially confounded with the variables representing the fixed time effects in the model. Effects of both sets of variables cannot be estimated simultaneously. Similarly, covariates that are too effective at explaining the random effects cannot be used because they induce a form of multicollinearity between the covariates and the variables representing the fixed time and time \times TRT effects in the model, and this may result in instability in the estimation of the corresponding TRT group means

Table 3. Repeated-measures ANOVA and SAS code for analyzing long-term experiment data assuming that covariates can explain the random effects of year and year \times TRT. For demonstration, two covariates, X1 and X2, are considered. Analysis assumes a randomized complete block design and incorporates models for the correlation structure. Variable names in the program are in capital letters.

Source†	df	SAS code for analysis‡
Block	$n - 1$	<pre>proc mixed method= reml data=sef1; class BLOCK TRT TIME; model Y = TRT X1 X2 X1*TRT X2*TRT TIME TRT*TIME / ddfm=kr hTYPE=1; random BLOCK; repeated / subject= BLOCK*TRT type=____; lsmeans TRT*TIME / diff at means; contrast ...; run;</pre>
TRT	$t - 1$	
Block \times TRT	$(n - 1)(t - 1)$	
Time	$r - 1$	
Time \times TRT	$(r - 1)(t - 1)$	
X1	1	
X2	1	
X1 \times TRT	$t - 1$	
X2 \times TRT	$t - 1$	
Error	$t(r - 1)(n - 1) - 2t$	
Total	$ntr - 1$	

† TRT, treatment.

‡ Options for "type=" are described in the text. Lsmeans and contrast statements vary according to research needs.

at each time. These means may be estimated wildly high or low as a result.

Recommendations

Ultimately, neither of these methods is guaranteed to work well for every problem. Not all response measurements can be reasonably approximated by known, simple models, and although explaining the random effects through covariates is appealing, implementation difficulties prevent it from being a viable alternative. The recommendation, then, is to use a model for the means if at all possible. Note that the two assumptions can be combined: one can incorporate into an analysis both a model for the means and covariates for the random effects. Regardless of the approach, nothing works exactly right, but ignoring the problem is clearly a worse option.

IMPROVED DESIGN AND ANALYSIS OF LTEs

As described earlier, the fundamental problem with the standard design of LTEs is the lack of replication of the sequence of times under which measurements are taken. Obviously, then, any experimental design that includes replication of time sequences is an improvement over the standard design.

Staggered-Start Design

Consider for a moment the situation with single-year experiments. It is typical that a single-year experiment is run in several years in order to confirm that the TRT effects do, indeed, hold up against varying environmental factors. Replication within a year is often done (for example, each year's experiment may be run in a randomized complete block design at a particular location), but for the purpose of estimating TRT effects and assessing their consistency across years, this replication is unnecessary. A combined analysis can be conducted using the TRT means within each year, and treating year as a block factor, to achieve these goals.

Ideally, then, replication in a LTE should come from running the experiment in several independent sequences of years. The duration of a LTE makes it impractical, however, to copy the single-year-experiment model and wait until the end of one replicate before starting another. This is especially true for LTEs whose duration is not specifically fixed—no one knows when the second replicate would begin!

As a compromise, a staggered-start design (Smith, 1979; Preece, 1986; McRae and Ryan, 1996; Martin et al., 1998; Orchard et al., 2000; Walters et al., 1988) is a practical alternative. This design, depicted in Fig. 5, staggers the start of each block of the experiment, so that successive replicates are established in successive years. In this way, each replicate experiences a different establishment-year environment, a different second-year environment, and so on. If there are n replicates, then measurements taken from a given time (i.e., growth seasons since establishment) are taken under n consecutive years, just like in a single-year experiment. Analyses

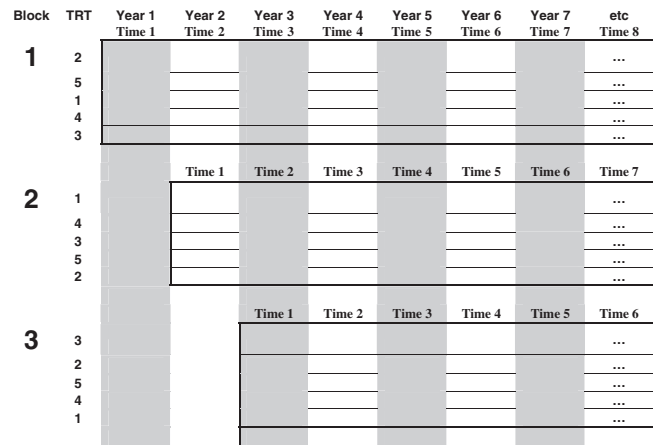


Fig. 5. Schematic of a staggered-start design for a long-term experiment.

of the results from any particular time are therefore valid in the *Agronomy Journal* sense of having been demonstrated under varying environmental regimes.

The measurements taken from different blocks in the same year do remain subject to random environmental effects associated with that year. The important difference between the staggered-start and the standard design is that those uncontrolled fluctuations affect measurements from different times, and hence can be measured separately from the fixed time effects. This makes the staggered-start design a statistically valid design for separating the fixed time and random year effects in a LTE, permitting the estimation of TRT means and standard errors at each time that are more representative of what might be expected to occur on average in a new run of the experiment.

Analysis Details for Staggered-Start Design

Although the staggered-start design has been proposed before for use in LTEs, a complete ANOVA for a staggered-start design has never been presented. The analysis of a staggered-start design must account for the fixed effects of TRT, time, and time \times TRT, as well as the random effects of rep, year, and various interactions of these factors with each other and with the fixed effects. The determination of which random-effect interactions to specify in the model must be made based upon careful consideration of all of the experimental units that are created by the structure of effects in the model.

This process is not uncommon; for example, the unique structure of AVOVAs for split-plot and strip-plot designs, as well as for extensions of these designs such as split-split-plot designs, arises from the identification of experimental units for each fixed effect in the experiment (see Milliken and Johnson, 1992; Mead, 1988; Kuehl, 2000). Once all experimental unit sizes and shapes are identified, then an error term for each fixed effect is created by combining the variability from all random effects that are measured on the same-sized experimental unit as the fixed effect.

In the present context, there are potential random effects associated with rep, year, rep \times year, rep \times TRT, year \times TRT, rep \times year \times TRT, REP \times time, year \times

time, $REP \times year \times time$, $rep \times TRT \times time$, $year \times TRT \times time$, and $rep \times year \times TRT \times time$. Examination of Fig. 5 is helpful in determining the experimental units upon which each effect is measured. Using the principle that an interaction among effects is measured on a unit corresponding to the intersection of those factors, the following results are revealed: (i) the $rep \times TRT$ random effect is the only one measured on the same unit as the TRT fixed effect and so it serves as the error term for TRT; (ii) the $rep \times year$, $rep \times time$, $year \times time$, and $rep \times year \times time$ random effects are all measured on the same unit as the time fixed effect, and so they get pooled together to form the error term for time; and (iii) the $year \times TRT$, $rep \times year \times TRT$, $rep \times TRT \times time$, $year \times TRT \times time$, and $rep \times year \times TRT \times time$ random effects are all measured on the same unit as the $time \times TRT$ fixed effect, and so they get pooled together to form the error term for $TRT \times time$. These rules lead to the ANOVA table shown in Table 4. Consideration of SAS's rules for pooling effects leads to the corresponding PROC MIXED code given in Table 4. Note that the LSMEANS and contrasts can be computed from this model in the same manner as with any typical design. Furthermore, the potential for serial correlation among measurements made on a given plot remains present in this design. Therefore, the process of modeling the correlation structure should be carried out as in the previous analyses.

To demonstrate the potential for substantial improvement in the quality of statistical analysis that results from the staggered-start design, the hypothetical example discussed earlier was reconsidered. The pattern of true means shown in Fig. 2a was retained, as were the random effects from the first 5 yr of the experiment. Because a staggered-start design for this same experiment would take 8 yr, additional comparably-sized random effects for year and $year \times TRT$ were added for the three additional years. Individual plot data were generated using the same procedures as before, resulting in a set of data that is depicted in Fig. 6. The chaotic tendencies of these data are due to the fact that at each time, a given TRT is exposed to four different sets of

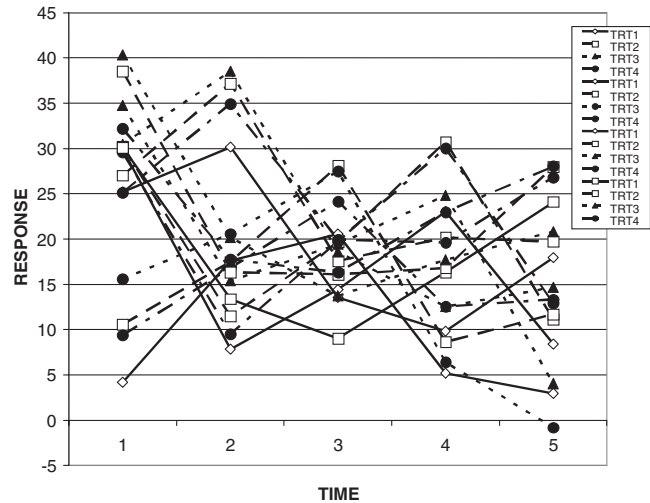


Fig. 6. Artificial example of data generated according to a hypothetical long-term experiment using the staggered-start design with four treatments in four blocks measured for 5 yr following establishment of plots.

random year and $year \times TRT$ effects (one for each replicate). An analysis was conducted according to the ANOVA in Table 4. Once again, different correlation structures were considered, and the AIC criterion indicated that compound symmetry was the best-fitting structure. Tests for TRT and $time \times TRT$ were highly significant ($P < 0.0001$) while the time main effect was not ($P > 0.20$). The means estimated by this model are shown in Fig. 7. There is a striking similarity between these means and the true means from Fig. 2a, much more than with any of the other analyses. Particularly, with the exception of one small reversal at Time 3, this is the only analysis among those considered that gets the means in the proper order at each time.

DISCUSSION

The data set used for the analysis of the staggered-start design was created to be comparable to the data set

Table 4. ANOVA and SAS code for analyzing staggered-start long-term experiment data. Analysis incorporates models for correlation structure. Variable names in the program are in capital letters.

Source†	df	SAS code for analysis‡
Block	$n - 1$	proc mixed method=
TRT	$t - 1$	reml data=set1;
Block \times TRT	$(n - 1)(t - 1)$	class BLOCK TRT TIME;
Year	$n + r - 3$	model Y = TRT TIME
Time	$r - 1$	TRT*TIME / ddfm=kr;
Block \times year	$nr - 2n - 2r + 4$	random BLOCK YEAR
\times time		BLOCK*YEAR*TIME;
Time \times TRT	$(r - 1)(t - 1)$	repeated / subject=
Error	$(n - 1)(t - 1)(r - 1)$	BLOCK*TRT type=_____;
Total	$ntr - 1$	lsmeans TRT*TIME / diff;
		contrast ...;
		run;

† TRT, treatment.

‡ Options for "type=" are described in the text. Lsmeans and contrast statements vary according to research needs.

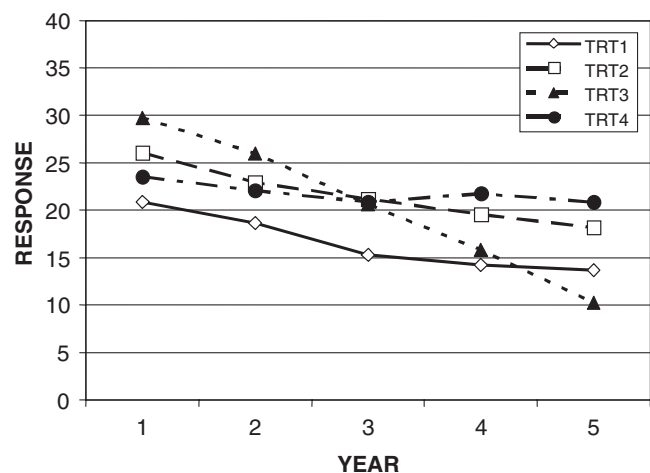


Fig. 7. Estimated treatment means at each time from the analysis of staggered-start data.

used for the analyses of the standard design with respect to the interference of the random effects. It is because of the staggering of the start of each successive block that the random effects of year and the fixed effects of time can be separated. The corresponding analysis is therefore better able to estimate TRT means than any of the possible analyses from the standard design. Although this is only a single example, the principle will hold any time the random environmental factors associated with each year may affect the measurements being taken in a LTE.

It should be noted that, in experiments of very long duration, such as the Rothamsted classicals, one might expect that the different TRT regimes eventually reach a steady state. It might be reasonable to assume, for example, that the true yield potential of a plot is not changing appreciably between Years 71 and 72 on TRT. If that happens, then there is a point after which there is no longer any practical amount of fixed change in the TRT means as time progresses, and only random effects remain. In that case, an analysis of average responses taken during a sufficiently long period of time on each plot (i.e., a summary statistics approach using a sample average summary) suffices to provide inference on the steady-state values of these TRT means. One should be aware, however, that subtle changes in TRT means over time cannot be detected in this manner, so this approach to analysis should be used only when it is fairly certain that an equilibrium has been reached.

Design Considerations

In ordinary experiments run in randomized complete block designs, it is sometimes arranged to have some or all TRTs randomized to more than one unit within a block. This is referred to as a generalized randomized complete block design (Steel and Torrie, 1980). This additional replication can be useful for ensuring that each TRT will, indeed, provide a response in each block, even if something happens to cause loss of data for a plot. It also can provide some measure of the consistency of TRT effects within blocks, which can be useful if blocking is based on a factor (such as initial fertility) that may interact with the TRTs being used. These considerations aside, however, within-block replication is not strictly necessary from a purely statistical point of view as long as the blocks can be reasonably considered to be random effects. Valid tests for the means are obtained by using the block \times TRT interaction as an error term. See Steel and Torrie (1980) for a discussion.

The same logic applies to blocks in a staggered-start design. There is no statistical reason requiring that there be replication within a block, meaning that there need not be more than one plot of a TRT initiated in a particular year. Certainly, a researcher may choose to run multiple plots of any TRTs within one block, particularly if there is some concern that plots may need to be abandoned before the end of the experiment for some reason. However, a nice feature of the staggered-start design is that it requires few additional resources beyond what might be planned for the standard design.

The space requirements are the same; one needs only delay initiating the experiment by one year on successive blocks, rather than starting them all at once. The staggering of the start does add one year to the duration of the experiment for every additional block beyond the first.

Split units and factorial TRT structures are easily incorporated into a staggered-start LTE. One simply designs the experiment as one would any other blocked experiment. The plot layout is exactly the same for a staggered-start design as it would be under the standard design. The only change is in staggering the establishment of successive blocks to take place in different years.

Analysis Considerations

As indicated by Table 4 and the discussion above, the analysis of a staggered-start LTE is not really any more difficult than the analysis of ordinary repeated-measures experiments. The procedures for modeling the correlation structure are the same and can be carried out using any mixed-model software that allows specification of correlation structures. M.J. Poehlman (2003, Design and analysis of long-term field trials with annual measurements; unpublished M.S. report, Dep. Statistics, Kansas State Univ., Manhattan) found that the tests for fixed effects for a staggered-start design maintain their designated type I error rates when no serial correlation is present. Research is presently underway to examine the performance of the ANOVA for staggered-start LTEs on data originating from a variety of different possible correlation structures. The use of a summary-statistics approach to analysis of a staggered-start LTE also needs to be studied more carefully to ensure that the presence of different levels of random effects on measurements taken at the same time-since-initiation does not adversely impact the quality of the analysis.

ACKNOWLEDGMENTS

Thanks are owed to Drs. Stephen Machado and Steve Petrie for the invitation to speak at the 2005 ASA-CSSA-SSSA meetings and for organizing this series of papers. I also wish to thank the faculty and students from the Department of Agronomy at Kansas State University for their patience in explaining their projects to me during the past 13 years.

REFERENCES

- Allredge, J.R., and F.L. Young. 1995. Issues in the Analysis of a Long-Term, Integrated Pest Management Field Study, p. 101–111. *In Proc. of the 7th Annual Kansas State Univ. Conf. on Applied Statistics in Agriculture*. Kansas State Univ., Manhattan.
- Aulakh, M.S., N.S. Pasricha, H.S. Baddesa, and G.S. Bahl. 1991. Long-term effects of rate and frequency of applied P on crop yields, plant available P, and recovery of fertilizer P in a peanut-wheat rotation. *Soil Sci.* 151:317–322.
- Bailey, T.B., J.B. Swan, R.L. Higgs, and W.H. Paulson. 1996. Long-term tillage effects on continuous corn yields, p. 18–32. *In Proc. of the 8th Annual Kansas State Univ. Conf. on Applied Statistics in Agriculture*. Kansas State Univ., Manhattan.
- Barber, S.A. 1979. Soil phosphorus after 25 years of cropping with five rates of phosphorus application. *Commun. Soil Sci. Plant Anal.* 10: 1459–1468.
- Boström, U., and H. Fogelfors. 2002. Long-term effects of herbicide

- application strategies on weeds and yield in spring-sown cereals. *Weed Sci.* 50:196–203.
- Cochran, W.G. 1939. Long-term agricultural experiments. *J. R. Stat. Soc. [Ser. A]* 6(Suppl.):104–148.
- Guerin, L., and W.W. Stroup. 2000. A simulation study to evaluate PROC MIXED analysis of repeated measures data. p. 170–203. *In Proc. 12th Kansas State Univ. Conf. on Applied Statistics in Agriculture.* Kansas State Univ., Manhattan.
- Johnston, A.E. 1994. The Rothamsted classical experiments. *In R.A. Leigh and A.E. Johnston (ed.) Long-term experiments in agricultural and ecological sciences.* CAB International, Wallingford, UK.
- Kuehl, R.O. 2000. Design of experiments: Statistical principles of research design and analysis. 2nd ed. Duxbury Press, North Scituate, MA.
- Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. 1996. SAS system for mixed models. SAS Inst., Cary, NC.
- Loughin, T.M., M.P. Roediger, G.A. Milliken, and J.P. Schmidt. 2006. On the analysis of long-term experiments. *J. R. Stat. Soc., Ser. A.* Published online 19 July 2006 at <http://www.blackwell-synergy.com/toc/rssa/0/0> (verified 28 Sept. 2006).
- Martin, B., J. Bennett, C. Cullis, D. Godwin, and W. Mason. 1998. Assessment of long-term experiments in Australia, Kingston, ACT, Australia. Grains Research and Development Corporation, Barton, ACT, Australia.
- McCollum, R.E. 1991. Buildup and decline in soil phosphorus: 30-year trends on a Typic Umbraquult. *Agron. J.* 83:77–85.
- McRae, K.B., and D.A.J. Ryan. 1996. Design and planning of long-term experiments. *Can. J. Plant Sci.* 76:595–601.
- Mead, R. 1988. The design of experiments: Statistical principles for practical application. Cambridge Univ. Press, Cambridge.
- Milliken, G.A., and D.E. Johnson. 1992. Analysis of messy data, Volume I: Designed experiments. Chapman and Hall, New York.
- Orchard, B.A., B.R. Cullis, N.E. Coombes, J.M. Virgona, and T. Klein. 2000. Grazing management studies within the temperate pasture sustainability key program: Experimental design and statistical analysis. *Aust. J. Exp. Agric.* 40:143–154.
- Ott, L., and M. Longnecker. 2001. An introduction to statistical methods and data analysis. 5th ed. Duxbury Press, North Scituate, MA.
- Patterson, H.D. 1964. Theory of cyclic rotation experiments. *J. R. Stat. Soc. Ser. B* 26:1–45.
- Poulton, P.R. 1996a. The Rothamsted long-term experiments: Are they still relevant? *Can. J. Plant Sci.* 76:559–571.
- Poulton, P.R. 1996b. Management and modification procedures for long-term field experiments. *Can. J. Plant Sci.* 76:587–594.
- Preece, D.A. 1986. Some general principles of crop rotation experiments. *Exp. Agric.* 22:187–198.
- SAS Institute. 2004. SAS 9.1.3 help and documentation. SAS Inst., Cary, NC.
- Schlegel, A.J., and J.L. Havlin. 1995. Corn response to long-term nitrogen and phosphorus fertilization. *J. Prod. Agric.* 8:181–185.
- Smith, A. 1979. Changes in botanical composition and yield in a long-term experiment. *In A.H. Charles and R.J. Haggard (ed.) Changes in sward composition and productivity.* Br. Grassland Soc., Reading, UK.
- Steel, R.G.D., and J.H. Torrie. 1980. Principles and procedures of statistics: A biometrical approach. McGraw Hill, New York.
- Walters, C.J., J.S. Collie, and T. Webb. 1988. Experimental designs for estimating transient responses to management disturbances. *Can. J. Fish. Aquat. Sci.* 45:530–538.
- Webster, R., and R.W. Payne. 2002. Analyzing repeated measurements in soil monitoring and experimentation. *Eur. J. Soil Sci.* 53:1–13.
- Yates, F. 1954. The analysis of experiments containing different crop rotations. *Biometrics* 10:324–346.