Published online May 9, 2019

CHAPTER 8: THE ANALYSIS OF COMBINED EXPERIMENTS

Philip M. Dixon,* Kenneth J. Moore, and Edzard van Santen

Agronomic experiments are often replicated over time (e.g., seasons or years), space (e.g., locations), or both to evaluate treatments across multiple environments; the subsequent statistical analysis is commonly called the analysis of combined experiments. The details of the analysis depend on the desired scope of inference and the model for treatment effects. Narrow-sense inferences are conclusions about the years, locations, and treatments included in the experiment. In a narrow-sense analysis, the environment × treatment interaction is modeled as a fixed effect. Broad sense inferences make conclusions about treatment effects in larger population of years and locations. In this analysis, the treatment × environment interaction is modeled as a random effect. Treatment effects can be modeled as a factor; this provides estimates of treatment means or linear contrasts among those means. Or, treatments can be considered as levels of a continuous variable; this estimates a response curve. In this chapter we provide guidance on how to choose among the various models and provide details for both narrow- and broad-sense inference. We use a meat triglyceride study and an oat (*Avena sativa* L.) cultivar trial to illustrate the analysis of group means and a fertilizer response trial to illustrate the analysis of a response curve. Finally, we discuss heterogeneity and pooling of variances.

Agronomic experiments are often repeated in multiple growing seasons and/or multiple geographic locations to expand the scope of the conclusions. Collectively, times, locations, or their combinations will be called environments. Experiments repeated in multiple environments come in two slightly different forms with slightly different goals. One form, which we will call an agronomic practice experiment, has the goal of understanding treatment effects across a range of environments. This provides information about both the treatment effects averaged over a broad range of environments and the consistency of effects over such environments (Blouin et al., 2011). Analysis of this form of study is often called the combined analysis of multiple experiments. The second form of multienvironment study, commonly done by plant breeders, evaluates the performance of many genotypes across multiple environments. This type of study often has the goal of understanding details of the interaction between geno-

Abbreviations: AIC, Akaike Information Criterion; EMS, Expected Mean Square; LR, Likelihood ratio; REML, Restricted maximum likelihood; RCBD, Randomized complete block design.

P.M. Dixon, Department of Statistics, Iowa State University, IA 50011; K.J. Moore, Department of Agronomy, Iowa State University, IA 50011; E. van Santen, IFAS Statistical Consulting Unit and Department of Agronomy, University of Florida, FL 32611. *Corresponding author (pdixon@iastate.edu)

doi:10.2134/appliedstatistics.2016.0004

Applied Statistics in Agricultural, Biological, and Environmental Sciences
Barry Glaz and Kathleen M. Yeater, editors

types and environments, for which specialized models, such as the AMMI model (see e.g., Gauch 2006), are frequently used. In this chapter, we only discuss agronomic practice experiments and the combined analysis of multiple experiments. We only consider analyses of the plot-level data from all environments and do not consider two-stage analyses (Piepho et al., 2012) or meta-analysis (Koricheva et al., 2013).

A combined analysis of multiple experiments extracts more information from a given body of collected data than can be gained from a series of experiment-specific analyses. For example, Thompson et al. (1993), describe the additional information provided by a combined analysis of 12 grazing studies conducted at nine locations over a period of 13 yr: "The mixed models procedure permitted estimation of the fixed effects of treatments over a broad inference space of future years and different tall fescue pastures over a wide geographic range, detected relationships that had not been apparent in the individual studies, such as the interactions between clover presence and E+ infestation levels, and provided a more coherent body of information than did the results obtained from each discrete study."

One of the first instances of analysis of a combined experiment, actually a multi-environment plant breeding trial, is described in an exchange of letters between the USDA geneticist R.F. Immer and R.A. Fisher about the analysis of a six-location nine-year barley yield trial (Wright, 2013). Fisher (1935, p. 211–215) discusses the concepts of a combined experiment. He specifically contrasts separate analyses for each location and the combined analysis of all locations. Cochran and Cox (1957) in Chapter 14 of their classic textbook on experimental design describe the analysis of combined experiments, mentioning three rationales for conducting an analysis of a series of experiments, viz. planned recommendations for an "extensive population in space and/or time", investigating the effect of "external conditions on some measurement", and testing interlaboratory agreement of an assay. Laird and Cady (1969) provided one of the first examples for the analysis of a large-scale combined experiment that involved 76 locations. Laird and Cady (1969) discussed the problem of arriving at the proper error term when conducting a combined analysis, foreshadowing comments by Binns et al. (1983) regarding the nature of random error components in combined experiments in their response to the seminal paper by McIntosh (1983).

McIntosh (1983), to our knowledge, was the first to pull together in an easily accessible form the information a researcher needs to conduct an analysis of experiments across environments. She describes the analyses of combined experiments based on the constraint that the effects of random interactions, for example treatment by year interactions, sum to zero over each level of a random factor, such as year. This assumption reflects a tradition in the analysis of fixed effects that, in most cases, is not appropriate for a random effect (Hocking 1985, p. 332–334). Moore and Dixon (2015) describe the analysis of combined experiments under the assumption that all random effects are mutually independent, which many, such as Hocking (1985, p 332) and McLean et al. (1991), consider more reasonable.

In 1983, virtually all researchers in agronomy and crop science, except for quantitative geneticists, were operating in a fixed-effects-only environment, that is, the only random effect term was the residual; all other effects in the model, even blocks, were fixed. This is not to imply that 'modern approaches' didn't exist, just that researchers were not utilizing them. Applied researchers in those fields were not aware of the contributions of Yates (1937) and Nelder (1965a, b) and if they had been, the computational resources needed would not have been available to them. The state of the art data analysis in the 1980s was described as "Making ordinary least squares the 'usual case'- essentially treating the mixed model as something of a curiosity of occasional importance– creates a number of difficulties in data analysis, some unrecognized or not fully appreciated, others ignored" (Stroup, 1989). At the heart of the matter, as Stroup (1989) pointed out, is that real-world error structures are more complex than the simple models assumed under fixed effects models. There were notable exceptions to the widespread use of fixed effects models, for example by researchers supported by the Scottish Agricultural Statistics Service after the implementation of restricted maximum likelihood (REML) in the 1970s (Robinson et al., 1982), but most agronomic researchers did not have this tool available to them.

Developments in statistical computing over the last 50 yr have enabled researchers to use mixed models procedures on a regular basis. Two key developments were the development of the Wilkinson algorithm for balanced multistratum problems (Wilkinson, 1970), implemented in the early 1970s in the GenStat package, and the REML algorithm (Patterson and Thompson, 1971), which enabled the analysis of unbalanced data. The credit for initiating a generally applicable software application in the United States goes to the Southern Regional Project S-189 entitled "Statistical computing methodology for research planning and analysis", a project of the University Statisticians of Southern Experiment Stations, commonly known as USSES, a group that since its inception in 1962 has had a successful career in providing information for appropriate statistical analyses of agronomic data. SAS PROC MIXED has its roots in the cooperation between USSES and the SAS Institute. Giesbrecht (1989) pointed out "The initial goal in the development of MIXMOD [the progenitor of PROC MIXED] was to create a relatively general tool that researchers could use to estimate variance components in unbalanced datasets. The emphasis has always been to preserve generality while achieving reasonable computational efficiency." The Statistics Department at Rothamsted Experiment Station in England played a similar role in the late 1960s with the development of the GenStat package. Subsequently, contributions such as (Piepho et al., 2003; Piepho et al., 2004; Piepho et al., 2006) have provided practical road maps for the use of mixed model methodology in the analysis of agronomic data.

It is understood that the preceding discussion should not be construed as a rigorous and exhaustive historical treatise of worldwide efforts in the development of practical approaches to mixed modeling. It only functions in setting the stage for the discussion following.

In this chapter, we describe two models commonly used for the analysis of combined experiments and elaborate on some practical details. These include deciding whether interactions should be fixed or random, choosing whether to subdivide the treatment by environment interaction, and using the data to decide whether or not to assume equal residual variances for all environments. The most important SAS code is provided in boxes in the paper; additional SAS code is in the Supplemental Material. R code is provided in the Supplemental Material.

A Linear Model for Qualitative Treatments in a Combined Experiment

We consider an experiment with qualitative treatments, for example different forms of tillage or weed management strategies, conducted in a randomized complete block design with each treatment occurring once per block. If this experiment is conducted in one environment, a model commonly used to evaluate differences among treatments is:

$$Y_{jk} = \mu + \beta_j + \tau_k + \varepsilon_{jk} \qquad [1]$$

where:

Y_{jk} is the response measured on the plot with treatment k in block j,

β_j is the effect of block j,

τ_k is the effect of treatment k, and

$\varepsilon_{\mathrm{jk}}$ is the experimental error for plot jk.

We use the common model for a randomized complete block design (RCBD), because RCBDs are very commonly used in agronomic investigations. The model in Eq. [1] can be altered in standard ways when some other field design (e.g., completely randomized, Latin square, lattice) is used.

Commonly, the goal of the study is to estimate and describe differences between treatment means or linear contrasts among treatment means. To simplify descriptions in the rest of the chapter, we refer to treatment differences, but everything said about treatment differences applies more generally and equally well to linear contrasts among treatment means. When treatment differences are the focus of a study, the overall mean and the treatment effects are modeled as fixed effects and the experimental errors are modeled as random effects. Blocks are sometimes modeled as random effects and sometimes modeled as fixed effects. The consequences of this choice are described by Dixon (2016) and summarized in the next section.

When this experiment is repeated in multiple environments, for example multiple years, multiple locations, or both, Eq. [1] is extended to account for variability among environments. A commonly used model for such an experiment is:

$$Y_{ijk} = \mu + E_i + \beta_{j(i)} + \tau_k + E\tau_{ik} + \varepsilon_{ijk} \qquad [2]$$

where:

Y_{ijk} = response measured on the plot with treatment k in block j in environment i,

μ = overall mean,

E_i = effect of environment i,

β_{ij} = effect of block j in environment i,

τ_k = effect of treatment k,

$E\tau_{ik}$ = interaction effect for treatment k in environment i, and

ε_{ijk} = experimental error for plot jk in environment i.

Again, the overall mean, μ, and the treatment effects, τ_k, are commonly modeled as fixed effects and the experimental error, ε_{ijk}, is modeled as a random effect. Each of the other terms in Eq. [2] may be modeled either as a fixed or random effect. The choice for environment and block effects can change some results, but often has no influence on the conclusions about treatment differences (see next section). However, the choice (fixed or random) for the environment by treatment interaction, $E\tau_{ik}$, is crucial, so that choice will be discussed separately.

The Choice of Fixed or Random Effect in General

Most of the terms in either Eq. [1], for a single-environment study, or Eq. [2], for a multi-environment study, may be modeled as fixed effects or random effects. Gelman (2005) summarizes five criteria that have been proposed to decide whether an effect should be modeled as fixed or as random. Of these, the most common is whether the effects in the model are a random sample from some probability distribution (e.g., Searle, 1971; p. 392; McCulloch, Searle and Neuhaus 2008, p. 16; Robinson 1991). Sometimes, the random sample part is dropped, for example, "do these effects come from a probability distribution"? (Robinson 1991). Sometimes, the criterion is elaborated, for example, "the context of the data, the manner in which they were gathered, and the environment from which they came" (Searle 1971, p. 382).

A clear answer to the question of fixed or random may not be available (McCulloch, Searle, and Neuhaus, 2008, p. 17). For example, consider a hypothetical agronomic field experiment. Because the study field is expected to be heterogeneous, and that heterogeneity is expected to be spatially organized, the field is divided into blocks of (usually) contiguous plots. To simplify the description, consider a study conducted in a 0.5 ha field divided into five blocks. Each block is 0.1 ha in size and contains four plots. Should block effects be modeled as a fixed effect or as a random effect? In the absence of a treatment, two plots in the same block are expected to be more similar to each other than are two plots in different blocks. This suggests a random effect model. In incomplete-block or row-column designs, block effects are almost always considered random to recover interblock information.

If the 0.1-ha blocks in our hypothetical study are considered as a random effect, it is not clear what population those random effects are sampled from. The blocks are either a complete enumeration of all five 0.1-ha parts of the 0.5-ha study field or they are a sample from some population of 0.1-ha size patches of land. If the latter, neither the population nor the mechanism used to draw the sample of blocks is known. The sampling issue is even more serious when blocks are chosen as internally homogeneous areas of a field, as suggested by Stroup (2013, p. 448). Similar issues apply to other sorts of blocks, such as groupings of animals based on initial weight (Dixon, 2016). Dixon (2016) argues that blocks should be considered a fixed effect, unless there are specific reasons to consider them random. Others (e.g., Casler, 2018 in Chapter 3) disagree.

For many study designs, the choice of random or fixed for block or environment effects changes statistical results about treatment means (Dixon, 2016) but has no consequence for statistical results about differences among treatment means (Giesbrecht and Gumpertz, 2004, p. 86). A very common study design with this property is the RCBD in which each treatment occurs once in each block. It is also true for generalizations where the columns of the design matrix for blocks are orthogonal to those for treatments. However, the interpretation of the statistical results is not the same. When block effects are considered fixed, the interpretation describes differences on these specific blocks. When blocks are considered random, the interpretation is about differences in the wider population. Although we refer frequently to differences among treatment means, all of our statements about those differences apply equally well to linear contrasts among treatment means.

The Choice of Fixed or Random Effect for the Environment by Treatment Interaction

The choice of fixed or random for the environment by treatment interaction effects, $E\tau_{ik}$, has major consequences for inferences on differences between treatment means (Newman et al., 1997). These consequences can be seen in two ways: by computing the variance of differences between two treatment means or by computing the Expected Mean Square (EMS) for the treatment effect (Table 1). The model used for these computations is Eq. [2]. To simplify the presentation, both experiments and blocks within experiments are considered fixed effects and we presume that the number of blocks, *J*, is the same in each of the *I* experiments. When blocks or environments and blocks are random, there are additional terms in the EMS or in the variance of a treatment difference, but the fundamental principles described here still hold. When the number of blocks is not the same in all experiments, the expressions for the EMSs and the variance of the treatment difference do not simplify as nicely as they do with equal numbers of blocks, but again, the fundamental principles still hold.

The quantities relevant to both the EMS and the variance of a treatment difference are the variance of the experimental errors, σ^2_{error}, and the variance of the interaction, σ^2_{ET}. If the environment by treatment interaction, $E\tau_{ik}$, is a random effect, the random values of that interaction effect are assumed to have the variance σ^2_{ET}, be mutually independent of each other, and be independent of the random experimental errors. That is, we do not impose sum-to-zero constraints, for example, for values of $E\tau_{ik}$, within a treatment.

When the environment by treatment interaction is considered fixed, the variance of a treatment difference depends only on the error variance and the total number of plots used for that treatment. Since there is one plot per treatment in each block of each environment, each treatment occurs on *J* plots in each experiment, so the total number of plots is *IJ*. For this study design, the only random effect in the Expected Mean Square for Treatments is the error variance, so the appropriate denominator for an *F* test of no differences among treatment means is the mean square that estimates σ^2_{error}, that is, MS error.

When the environment by treatment interaction is considered random, the variance of a treatment difference includes a term that depends on the size of the environment by treatment interaction. In Eq. [2], the environment by treatment effect is considered to have one value for each combination of environment and treatment. That value is shared by all *J* plots for one treatment in one environment. The treatment mean is an average over *IJ* plots but only *I* environment by treatment effects, so the contribution of the environment by treatment interaction to the variance of the treatment mean is $\sigma^2_{\text{ET}}\ /I$. From Table 1, we see that considering the environment by treatment interaction as a random effect increases the variance of the treatment difference. The magnitude of that increase depends on the size of the environment by treatment interaction, quantified by σ^2_{ET}, and the number of environments in the study.

The choice of model for the environment by treatment interaction corresponds to the choice of an inference space (McLean et al., 1991). McLean et al. (1991) distinguish between three inference spaces: narrow sense, intermediate sense, and broad sense. In a multienvironment study, narrow-sense conclusions describe average treatment effects in the environments used in the study. To obtain these, the environment by treatment interaction is modeled as a fixed effect. Broad-sense conclusions describe average treatment effects in new environments. These new environments are assumed to be a random sample from a population of potential environments. To obtain broad-sense inferences, environments, blocks, and the environment by treatment interaction are modeled as random effects. The environment by treatment interaction quantifies the consistency of treatment effects across multiple environments. The precision of broad-sense conclusions about average treatment effects depends on the magnitude of the environment by treatment interaction; the precision of narrow-sense conclusions does not.

McLean et al. (1991) use intermediate-sense inference to describe conclusions from a model where the main effect of environment is fixed but the environment by treatment interaction is random. In a balanced design, the difference between intermediate- and broad-sense does not influence the variance of the treatment difference. It only influences the variance of the treatment mean. Because the focus of an agronomic practice experiment is on the difference between treatments, we will use broad-sense inference as defined by McLean et al. (1991) in our examples. Modeling environments as a fixed effect, so using intermediate-sense inference, will lead to different results for treatment means but the same, or very similar, results for treatment differences.

Example: Triglyceride Measurements in Meat

These points are illustrated by a small repeated experiment on two methods to measure triglyceride concentrations in meat. The data are in the Supplemental Material. On day one of the study, four pieces of meat are randomly assigned to one of two methods for chemical extraction and processing. Each piece of meat is measured once resulting in two measurements of triglyceride concentration for each method. This is repeated for a total of four days, giving a total of 16 observations. We are considering days as independent repetitions of a basic experiment; one single day of the study is a valid experiment, with replicated randomization of treatments to pieces of meat. Although you could consider days to be blocks, we prefer to consider them as separate environments, because data for each day could be analyzed separately.

TABLE 1. Consequences of the choice of fixed effect or random effect for the environment by treatment interaction. The variance of the experimental errors is $\sigma^2_{\text{error}}.Q(T_k)$ indicates a quadratic function of the treatment effects. When the environment by treatment interaction is random, the random values of that interaction effect are assumed to have variance σ^2_{ET} . Sample sizes are *I*: number of environments and *J*: number of blocks per environment.

	Fixed	Random
Variance of a treatment difference, $\bar{Y}_{..k}-\bar{Y}_{..l}$, averaged over I environments and J plots	$\frac{2\sigma^2_{error}}{IJ}$	$\frac{2\sigma^2_{ET}}{I}+\frac{2\sigma^2_{error}}{IJ}$
Expected Mean Square for Treatments	$Q(T_k)+\sigma^2_{\text{error}}$	$Q(T_k)+J\sigma^2_{\text{ET}}+\sigma^2_{\text{error}}$
Denominator MS to test H_0: $Q(T_k)=0$	MS Error	MS Expt × Trt

TABLE 2. ANOVA table for the triglyceride measurement data, with Expected Mean Squares (E MS) when day and the day*method interaction are considered a fixed effect and when both are considered a random effect. In both cases, the main effect of treatment is considered a fixed effect. The Q() notation indicates a quadratic function of the specified model coefficients.

Source	Degrees of freedom	Mean square	E MS: fixed	E MS: random
day	3	143.8	$\sigma_e^2+Q(d,dm)$	$\sigma_e^2+2\sigma_{dm}^2+4\sigma_d^2$
method	1	406.0	$\sigma_e^2+Q(m,dm)$	$\sigma_e^2+2\sigma_{dm}^2+Q(m)$
day*method	3	36.1	$\sigma_e^2+Q(dm)$	$\sigma_e^2+2\sigma_{dm}^2$
error	8	14.4	σ_e^2	σ_e^2

Within each day, there are no additional restrictions on randomization, so an appropriate model (Eq. [3]) is a simplification of Eq. [2] without block effects.

$$Y_{ijk}=\mu+d_i+m_j+dm_{ij}+\varepsilon_{ijk} \qquad [3]$$

where d_i is the effect of day *i*, m_j is the effect of method *j*, and dm_{ij} is the interaction effect for day *i* and method *j*. The random error, ε_{jik}, is associated with a single measurement on an individual piece of meat.

A dot plot of the data (Fig. 1) shows the major patterns in the data:

1) the average measurement on Day 2 is lower than the average on other days,
2) the two methods have approximately similar measurements on Day 1,
3) on subsequent days, measurements for method A are substantially larger than those for method B, and
4) the average difference between the two methods is largest on Days 2 and 3.

The partial ANOVA table for these data, with degrees of freedom and Mean Squares, is given in Table 2. Table 2 also includes the Expected Mean Squares when the interaction is modeled as a fixed effect and when it is modeled as a random effect.

If we want to make conclusions about the average difference between methods A and B on these 4 d, we should use narrow-sense inference, so the day × method interaction is modeled as a fixed effect. The appropriate *F* test for the difference between the two methods is the ratio *F* = MS(method)/MS(error) = 28.16, which has a central *F* distribution with 1, 8 degrees of freedom under the null hypothesis of no difference between the two methods. This gives a *p*-value for the test of that null hypothesis of 0.0007. Alternatively, the standard error of the mean difference between the two methods is estimated as $\sqrt{2\left[\frac{14.4}{8}\right]}=\sqrt{3.60}=1.90$. A confidence interval for the mean difference will use a *t* statistic with the error degrees of freedom, here 8. The resulting 95% confidence interval for the mean difference is (5.70, 14.45). The SAS code to produce these results is in Box 1. R code to fit the same model is in the Supplemental Material.

We might instead believe that the true difference between the two treatments could differ from day to day, for many subject-specific reasons. Narrow-sense inference describes the treatment effects for the four days used in the study, but when there is day-to-day variation in the treatment difference, the results are unlikely to be repeatable on a new set of four days. Broad-sense inference makes conclusions about the average treatment difference on four new days, assumed to be randomly chosen from a population of potential study days. The appropriate *F* test for the difference between the two methods is the ratio *F* = MS(method)/MS(day*method), based on inspection of the expected mean squares in Table 2. Under the null hypothesis of no difference, this ratio has a central *F* distribution with 1, 3 degrees of freedom because the mean square for the interaction has three degrees of freedom. The *F* statistic is 11.25, which gives a *p*-value of 0.044. A confidence interval for the mean difference will use a *t* statistic with the degrees of freedom (here 3). The resulting 95% confidence interval for the mean difference is (0.52, 19.63).

SAS code to compute the broad-sense results is in Box 2. To use this code, replace the proc mixed step in Box 1 with the code in Box 2. The SAS code code in Box 2 uses proc mixed with REML estimates of the variance components (the default) and the

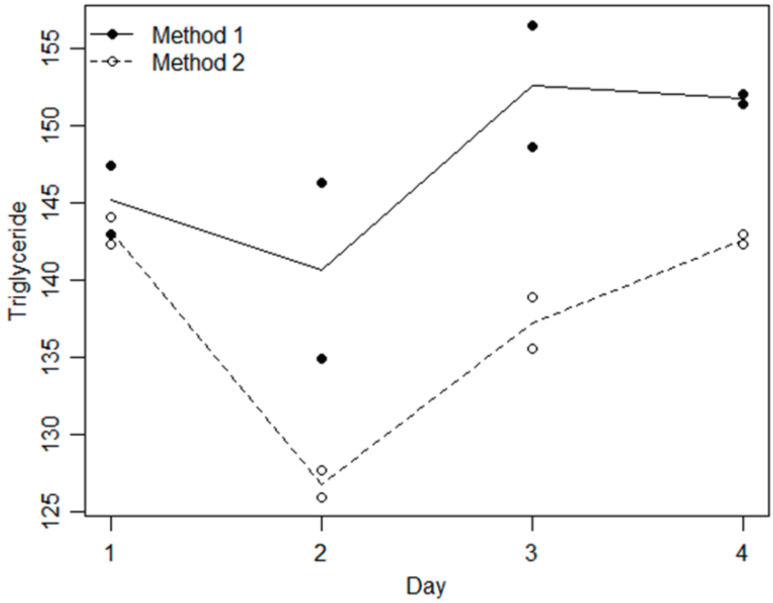

FIG. 1. Plot of the measured triglyceride concentration, as measured by two methods (A or B) on four days. Lines connect method-specific averages across the days.

BOX 1.

```
data trigly;
  infile 'triglybook.txt' firstobs=2;
  input method day y;
  run;

proc sgplot data=trigly;
  scatter x=day y=y /group=method;
  run;

proc mixed method=type3 data=trigly;
  class day method;
  model y = day method method*day;
  lsmeans method / diff cl;
  title 'Triglyceride example: Narrow sense inference';
  run;
```

BOX 2.

```
proc mixed data=trigly;
  class method day;
  model y = method /ddfm=satterth;
  random day method*day;
  lsmeans method /diff cl;
  title 'Triglyceride example: Broad sense inference';
  run;
```

TABLE 3. Comparison of broad- and narrow-sense inference for the mean difference between two measurement methods.

Inference about	Inference space	Number of days	Estimated difference	Standard error	Degrees of freedom	95% confidence interval
The 4 days in the study	narrow	4	10.08	1.90	8	(5.70, 14.45)
Four randomly chosen new days	broad	4	10.08	3.00	3	(0.52, 19.63)

generalized Satterthwaite approximation (Giesbrecht and Burns, 1985) for degrees of freedom. For this data set and design, other analysis choices, including method-of-moments estimates of the variance components (proc mixed method = type3), Kenward–Roger degrees of freedom adjustment, or use of proc glimmix, give identical results.

Table 3 compares the results from narrow- and broad-sense inference. The estimated differences are the same, but the broad-sense standard error of the difference is much larger. This reflects the additional uncertainty arising from apparent day-to-day variation in the treatment difference. The broad-sense 95% confidence interval is much wider because of both the larger standard error and the smaller degrees of freedom.

Should you use broad- or narrow-sense inference? The change from narrow- to broad-sense inference is the change from making a conclusion about the four days used in the study to making a conclusion about four new days randomly chosen from an estimated population of days. If we believe that the true difference between the two treatments on day *j*, that is, $(m_A + \mathrm{dm}_{\mathrm{Aj}}) - (m_B + \mathrm{dm}_{\mathrm{Bj}})$, is the same for each day, then the choice of days on which measurements are made is irrelevant. Narrow-sense conclusions are sufficient to describe the treatment difference. In fact, repeating the experiment on multiple days is unnecessary, except perhaps for practical reasons. If you want eight replicates of each treatment, but you can only process four measurements in a day, or a field only has room for four plots, you need to use multiple days or multiple fields.

When you have replicates on the same day or from the same field, it is possible to test for a non-zero interaction. For the meat data, the *F* test of the day by method interaction gives an *F* statistic of 2.5 and a *p*-value of 0.13. However, we suggest this test not be used to decide on narrow- or broad-sense inference because the test is answering the wrong question. A nonsignificant interaction does not imply that there is no interaction (Cox, 1958, p. 103). Claiming that the true treatment difference is the same each day is trying to use a statistical test to demonstrate the null hypothesis. Interaction tests in a factorial ANOVA have much lower power than do tests of main effects (Pearce, 1988). The power to detect a non-zero interaction can be increased by using a relaxed α level, such as 0.20, as suggested by Sulc et al. (2001)

and Edwards (1985, p. 263). This doesn't change the fundamental logical problem: that you are trying to prove the null hypothesis.

We suggest that the choice of broad- or narrow-sense inference be made based on study goals, not on characteristics of the data. Are you interested in describing what happened on the specific days used in the study? If so, narrow-sense inference is appropriate. The uncertainty in your results describes what could have happened if treatments had been assigned differently to the various replicates. Or, are you interested in describing what might happen if the study is repeated on new days? Now, broad-sense inference is appropriate. The uncertainty in your results describes what could have happened if treatments had been assigned differently to the various replicates *and* the study had been performed on a new set of days. This approach makes intuitive sense, but it calls for thinking about the goal(s) of the study at hand, not blindly following other examples.

Example: Oat Cultivar Trial

A more involved example is the analysis of an oat cultivar trial conducted in two years at three locations in Iowa. This study compared harvest index (HI) of 10 oat cultivars. Since the study goal is to compare mean HI among these cultivars, cultivar will be considered as a fixed effect. In each year and location, the experiment was conducted in a randomized complete block design with three blocks. There were no missing observations. This study is repeated in both space (Location) and time (Year). The analysis described in this section will consider the six combinations of location and year as six separate environments. The discussion of pooling in the next section describes other models for locations and years. Figure 2 shows the average Harvest Index for three cultivars in each of the six environments.

The initial analyses use Model 2 (Eq. [2]), with additive effects for environment, block within environment, and treatment, an interaction between environment and treatment, and an additive residual error. Because each cultivar occurs only once in each block, the plot-to-plot variation and the interaction between blocks and cultivars are combined into the error variation. This is a classic example of confounding two effects; we know both are present but they cannot be separately estimated from these data.

When narrow-sense inference, that is, inference about cultivar differences averaged over the six environments, is desired, the analysis requires only one random component, the residual error. Blocks can be included as a random effect if desired, but statistical results about treatment differences are unchanged because the design is orthogonal. The *F* test for cultivars has the error MS as the denominator; confidence intervals for differences or linear combinations among cultivar means use the error variance. SAS code to read and plot the data and fit the model for narrow-sense inference is in Box 3. In this code, blocks are considered a random effect nested in environments (the Environment variable); method = type3 is used to obtain the Mean Squares for the ANOVA table. Using REML estimates of the variance components provides the same *F* tests and *p*-values so long as the nobound option is used to allow a negative estimate of the block variance. The three estimate statements illustrate estimates that are often of interest: a cultivar mean averaged over environments, a difference between two cultivar means averaged over environments, and a difference between two cultivar means in a single environment.

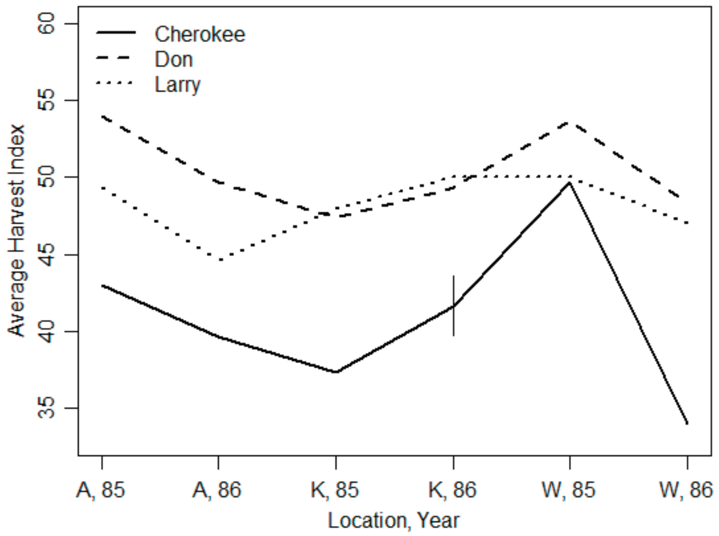

FIG. 2. Harvest index means for three varieties at each location and year, demonstrating the variety × environment interaction. Vertical bar shows + / - 1 se for the average harvest index for one variety and environment.

BOX 3.

```
proc import datafile=’oats.csv’ out=oats0 replace;
  guessingrows = 100;
  run;

data oats;
  set oats0;
  Environment = trim(Location) || ‘, ‘ || put(Year, 4.);
  run;

proc sort data=oats;
  by Environment;
  run;

proc sgplot data=oats;
  scatter x=Cultivar y=HI / group = Environment;
  title ;
  run;

proc mixed method=type3 data=oats;
  class Environment Block Cultivar;
  model HI = Environment Cultivar Environment*Cultivar /ddfm=kr;
  random Block(Environment);
  estimate ‘Cultivar: Cherokee’ intercept 1 Cultivar 1;
  estimate ‘Cultivar: Cherokee - Don’ Cultivar 1 -1;
  estimate ‘Cult: Cherokee - Don in Ames 1985’ Cultivar 1 -1
Environment*Cultivar 1 -1;
  lsmeans Environment*Cultivar /slice = Cultivar slice=Environment;
  title ‘Oat HI: Narrow sense inference, random blocks’;
  run;
```

Results of this analysis indicate a significant interaction between environment (year and location) and cultivar (Table 4). Two ways to understand this interaction are to plot the cultivar means in each environment and testing differences among cultivars in each environment. A plot of average Harvest Index for three cultivars across environments shows that some cultivars have similar means in all six environments, while other cultivars vary considerably across the environments (Fig. 2). This is confirmed by testing differences among environments within each cultivar. In SAS, this is accomplished by slicing the interaction by cultivar (Environment*Cultivar/slice = Cultivar). The table of cell means (i.e., means for each combination of environment and cultivar) is "sliced" into cultivar-specific components. The result is a collection of *F* tests, one for each cultivar, that evaluate differences among environments for that cultivar. Some cultivars (e.g., Larry) showed no evidence of difference in mean HI over the six environments, some (e.g., Don) showed weak evidence of difference, and others (e.g., Cherokee) performed differently in at least one environment (Table 5).

Testing cultivar-specific effects by slicing an interaction is similar, but not identical, to splitting the data into pieces, one for each cultivar, then testing differences among environments for each cultivar. It is not identical because the "slicing" *F* tests use pooled estimates of the error and block variances while the "splitting" tests use cultivar-specific estimates. The environment by cultivar interaction can be sliced either by cultivar or by environment. Slicing the interaction by environment, to

TABLE 4. ANOVA table for the narrow sense analysis of the oat variety trial example.

Source	Degrees of freedom	Mean Square	*F* statistic	*P*-value
Environment	5	294.27	33.65	< 0.001
Block(Environment)	12	8.74		
Variety	9	171.92	14.83	< 0.0001
Variety*Environment	45	26.57	2.29	0.0003
Error	108	11.59		

TABLE 5. *P*-values for slices of the cultivar by environment interaction. Each cultivar-specific test compares the average responses of a cultivar across six environments; each test has five numerator degrees of freedom. Each environment-specific test compares the average response of 10 cultivars in the specified environment; each test has nine numerator degrees of freedom.

Variety-specific tests		Environment-specific tests	
Variety	*P*-value	Location, Year	*P*-value
Cherokee	< 0.0001	Ames, 1985	0.0012
Don	0.074	Ames, 1986	0.0003
E77	0.0033	Kanawha, 1985	< 0.0001
Hytest	0.0064	Kanawha, 1986	0.0060
Lang	0.099	Washington, 1985	0.030
Larry	0.34	Washington, 1986	< 0.0001
Ogle	0.094		
Porter	0.0006		
Proat	< 0.0001		
Richland	< 0.0001		

BOX 4.

```
proc mixed nobound data=oats;
  class Environment Block Cultivar;
  model HI = Cultivar /ddfm=kr;
  random Environment Block(Environment) Cultivar*Environment;
  estimate 'Cultivar: Cherokee' intercept 1 Cultivar 1;
  estimate 'Cultivar: Cherokee - Don' Cultivar 1 -1;
  estimate 'Cultiver: Cherokee - Don in Ames 1985' Cultivar 1 -1 |
    Cultivar*Environment 1 -1;
  title 'Oat HI: Broad sense, 6 environments, REML variance
component
    ests';
  run;
```

produce *F* tests of cultivar differences within each environment, indicates that there was a significant cultivar response in each of the six environments (Table 5).

When broad-sense inference, that is, inferences about the differences among cultivars in new environments, is desired, the interaction between treatments and environments is considered a random effect. In addition, for broad-sense inference as defined by McLean et al. (1991), environment effects are considered random. Because a fixed effect cannot be nested inside a random effect, then block within environment must be a random effect. SAS code for broad-sense inference across 6 environments is in Box 4. Notice that the estimate statement to compute the difference between two cultivars in a specific environment requires a vertical bar (| symbol) to separate the fixed effects (cultivar) and the random effects (Cultivar × Environment).

As with the meat triglyceride example, the choice of narrow- or broad-sense inference has no effect on the estimated mean for one cultivar, averaged across environments, or the estimated mean difference between two cultivars, averaged across environments, but it does change the standard errors (Table 6). The standard errors are larger when using broad-sense inference because of variation between environments in both cultivar means and differences. However, broad-sense inference provides an estimate of the average performance of the cultivars over a population of environments.

The choice of narrow- or broad-sense inference does change the reported environment-specific difference between two cultivars (Table 6). This is because the narrow-sense value is an estimate of that difference, which involves two estimated interaction components. Under broad-sense inference, the interaction is a random effect, so components of that interaction are predicted using best linear unbiased predictions (BLUPs). In general, BLUPs are shrunk toward zero, so the narrow-sense estimates and broad-sense BLUPs are not the same (Robinson 1991). Hence, the broad-sense prediction is not the same as the narrow-sense estimate of an environment-specific difference.

Inferences using the broad-sense approach are to a larger population of environments than used in the actual study. Whether that population is hypothetical or real depends on how environments were chosen for the study. The population for broad-sense inference is clearly identifiable when the environments used in a study are a simple random sample from that population. For example, when six farms are randomly selected from a list of all farms in Iowa, the population for broad-sense inference is all farms in Iowa. If the environments are not chosen by a simple random sample, then broad-sense conclusions apply to a hypothetical population defined by the environments actually used in a study. Whether this hypothetical population reflects any real population depends on non-statistical information. In the meat example, the days used in the study are not a random sample from some population of days, but they could be imagined to represent a random sample from a hypothetical population of all the things that differ between days (e.g., lab conditions, properties of the piece of meat used that day).

The usual mixed models approach assumes infinite-size populations of random effects. Applied to a combined experiment, this means that the broad-sense variance calculations for treatment means or differences assume inference about an infinite population of environments. When the population of environments is not much larger than the number in the study, the calculated variances are too large because of the finite population correction factor (Thompson 1992, p. 15). Equations for variances of treatment differences and expected means squares for a finite population of environments are given in Bennett and Franklin (1954, Chapter 7). In the oat cultivar trial, the calculated variances are appropriate for inferences to a very large collection of locations and years. If the desired inferences are to five research farms in central Iowa, three of which were used in the study, the calculated variance is too large. In the extreme, if the entire population of environments were included in the study, after correctly adjusting for the finite population of environments, the variance of a treatment difference and the Expected Mean Square for treatments are the expressions in the Fixed column of Table 1. Again, this makes intuitive sense as you are in "possession" of all information regarding the target environments.

TABLE 6. Estimates and standard errors for the mean of one variety (Cherokee), a difference of two variety means (Cherokee – Don), and an environment-specific difference of two variety means (Cherokee – Don in Ames, 1995) for narrow- and broad-sense inference using six environments.

Quantity	Estimate for:		Standard Error for:	
	Narrow	Broad	Narrow	Broad
Mean for one variety	40.89	40.89	0.79	1.72
Difference of two varieties	-9.50	-9.50	1.13	1.72
Difference of two varieties in one environment	-11.00	-10.35	2.78	2.22

Choosing Whether to Subdivide the Treatment by Environment Interaction by Environment

The previous analysis of the oat cultivar data considered each combination of year and location as one of six unique environments. This assumes that the random effect associated with one year and location, for example, "Ames 1985", is independent of the random effects associated with any other year and location. Because all observations from a specific environment share one value of the

TABLE 7. Estimates, standard errors, and error degrees of freedom (df) for the mean of one variety (Cherokee), a difference of two variety means (Cherokee–Don), and an environment-specific difference of two variety means (Cherokee – Don in Ames, 1995) after subdividing environments into year, location and the year by location interaction. Degrees of freedom computed using the Kenward–Roger approximation.

Quantity	Estimate	Standard Error	Degrees of freedom
Mean for one variety	40.89	2.13	1
Difference of two varieties	-9.50	1.60	3.72
Difference of two varieties in one environment	-10.45	2.22	123

BOX 5.

```
proc mixed nobound data=oats;
  class Year Location Block Cultivar;
  model HI = Cultivar /ddfm=kr;
  random Year Location Year*Location Block(Year*Location)
    Cultivar*Year Cultivar*Location Cultivar*Year*Location;
  estimate ‘Cultivar: Cherokee’ intercept 1 Cultivar 1;
  estimate ‘Cultivar: Cherokee - Don’ Cultivar 1 -1;
  estimate ‘Cult: Cherokee - Don in Ames 1985’ Cultivar 1 -1 |
    Cultivar*Year 1 -1 Cultivar*Location 1 -1
    Cultivar*Year*Location 1 -1;
  title ‘Broad sense: subdivided interaction’;
  run;
```

environment random effect, a random effect for environment induces a correlation between observations from the same environment. Observations from different environments are independent, because they do not share any random variables. Six independent environments represent one of many possible ways to structure the random effects in the model. Other specifications of the random effects provide different models for the correlation among observations.

One alternative is to define three random effects: one for years (shared by all observations in the same year), one for locations (shared by all observations in the same location) and one for the interaction of year and location (shared only by observations in the same year and same location). This adds additional sources of correlation between observations. The model with random effects for year, location, and year × location allows observations from the same year to have one non-zero correlation, observations from the same location to have a different non-zero correlation, and observations from the same environment (year and location) to have a third non-zero correlation.

The same concept can be applied to subdividing a treatment by environment interaction into multiple components. For the oat cultivar data, the cultivar by environment interaction could be subdivided into three components: a cultivar by year interaction, a cultivar by location interaction and a cultivar by year by location interaction. Since the treatment by environment interaction represents the consistency of treatment differences across environments, subdividing the interaction allows one to assess consistency across years and a different consistency across locations.

SAS code to fit this model is in Box 5. Again, the nobound option is specified to allow variance components to be negative. Although a variance is strictly nonnegative, when viewed as a parameter determining the correlation between pairs of observations, a negative variance component specifies a negative correlation, which is entirely legitimate.

Selected results are given in Table 7. The estimated mean for one cultivar and estimated difference of two cultivars are unchanged by splitting environments, but the standard errors are different from those given in Table 6. Some standard errors increase compared with their “six environment” values while others decrease. Because there is less information about each component of the interaction, the degrees of freedom for estimated means and estimated differences are reduced when environment is split into year, location, and the year by location components. The Kenward–Roger (Kenward and Roger, 1997) approximate degrees of freedom for a cultivar mean and difference are 1 and 3.72 (Table 7), compared with 15.3 and 45 when environment is not split.

Choosing Whether to Subdivide the Treatment by Environment Interaction by Treatment

The pooled treatment by environment interaction combines information across treatments as well as across environments. So, the interaction could be subdivided by treatment instead of by environment. If the 10 cultivars represent two breeding groups and five maturity groups, the set of 10 treatments could be divided into a 2 × 5 factorial combination of breeding group × maturity group. Then, you could consider three components of the environment by treatment interaction: breeding group by environment, maturity group by environment, and the three-way interaction. Each of these contributions represents the consistency of that set of treatment effects across environments. A variance component close to 0 for one component means that those treatment contrasts are relatively consistent across the environments. A large variance component means that treatment contrasts are quite different across the environments. For example, if the breeding group by environment interaction variance is large, the difference between the two breeding groups varies across environments. In the usual analysis of a combined experiment, each of these interaction components is a random effect and would serve as the error term for the associated treatment effect. For example, breeding group by environment would be the error term for testing the breeding group effect.

Further subdivision of interaction components is possible. The differences among five maturity groups are described by four orthogonal contrasts. The environment by maturity group interaction could be subdivided into contributions from each orthogonal contrast. For example, if the difference between Maturity Group 1 and the other four maturity groups varied considerably across locations, but differences among the other four maturity groups were relatively consistent then the single degree of freedom contrast between Maturity Group 1 and the rest should be separated from the other three df among maturity groups. Such an analysis would include two different interactions, one for each group of contrasts.

Even further subdivision of interaction components is possible. The differences among the ten cultivars are described by nine orthogonal contrasts. Each could be assumed to have a different interaction with environment. The resulting model would have nine variance components for the environment by treatment interaction, one for each orthogonal contrast. When each orthogonal contrast is analyzed separately, this approach is known as within subject contrasts analysis (see for example, Keppel and Wickens, 2004, p. 358–359) or derived variables analysis (Diggle et al. 2002, p. 17).

A classic example of treatment-specific interaction components is the experiment on early lentils in Syria described by Peterson (1994, p. 207–215). The treatment structure is a three-way complete factorial with two levels of fertilizer, two levels of weevil control, and two levels of weed control. This was repeated at eight locations. Peterson's model included eight location interactions, one for each component of the treatment structure. Each location interaction is estimated with seven degrees of freedom. The location by fertilizer interactions are all small relative to the location by weevil and location by weed interactions (Peterson, 1994, p. 213). Peterson speculates that the eight locations had different levels of infestation by weevils and weeds. The response to the weevil treatment or the weed control treatment would be large at locations with heavy infestation and small at locations with little infestation. In contrast, the data suggests that the response to fertilization was similar at all locations. An alternative to Peterson's model would have three components to the interaction: location × weevil, location × weed, and a pooled location × other term. The location × other term would have 35 degrees of freedom. If the assumption of equal contributions to that interaction is appropriate, the 35 df estimate is a more precise estimate than any of the seven degree of freedom estimates.

The decision to pool or not to pool should preferably be made in the planning stage of a study, using the experimenter's understanding of sources and magnitudes of variability. This will be especially important when each individual experiment has multiple sources of variation; for example, variation among main plots and variation among split plots. A narrow-sense analysis will pool the main-plot variability across environments and separately pool the split-plot variability. The appropriate pooling for a broad-sense analysis is less clear. Is the consistency across environments of the main-plot treatment different from the consistency of the split-plot treatment? If so, interaction components should not be pooled. If the two consistencies are expected to be similar, interaction components could be pooled (Carmer et al., 1969).

A second consideration in deciding whether or not to pool is the error df for important tests (Edwards, 1985, p. 263). Increasing the error df greatly reduces upper quantiles of t and F distributions when the error df is small. For example, the 0.975 quantiles of t distributions are approximately 4.3 for two degrees freedom, 2.6 for five degrees freedom, 2.2 for 10 df and 2.1 for 15 df. Increasing the error df has a much smaller effect when the error df exceeds 20 or 30. For example, the 0.975 quantiles of t distributions are approximately 2.08 for 20 df, 2.04 for 30 df, 2.01 for 50 df and 1.98 for 100 df. Because of this characteristic, Edwards (1985, p. 263) suggests never pooling effects that have more than 20 or 30 df. He also suggests designing a study so that it has sufficient replication to provide 30, or at least 20, error df.

Careful consideration of both sources of variability and appropriate amounts of pooling is especially important in studies with multifactor treatment structures repeated in different types of environments. When both treatments and environments are structured, an analysis might subdivide both and consider all combinations of contributions. One author (PMD) has seen a combined experiment with a 2 × 2 × 2 treatment design repeated in two locations and two years. The proposed analysis subdivided environment into location, year, and the location × year interaction and subdivided treatments into 7 single degree of freedom contrasts. The analysis had over 20 treatment by environment interaction terms, (e.g., year × A, or location × B), each with one degree of freedom. This is the ultimate in subdivision and leads to tests with very low power because denominator degrees of freedom will be close to one for each test. The practical analysis is to pool interaction terms, using subject matter knowledge to decide which terms should be similar.

Choosing Whether or Not to Assume Equal Residual Variances for All Environments

The analyses of both the meat and the oat cultivar data sets assumed that all observations had the same error variance. Because each repetition of a repeated study is a stand-alone experiment, an error variance can be estimated separately for each repetition. When each repetition can be assumed to have the same error variance, these repetition-specific estimates are pooled to estimate a single common variance. Is it reasonable to use a single pooled variance, or should each repetition be allowed to have a different error variance?

One way to answer that question relies on an experimenter's knowledge of the study sites. The error variance describes the variability between experimental plots treated alike. When the repetitions are at the same location, that is, repeated in different years, it might be more reasonable to use the pooled estimate. When a study is repeated at different locations, "common sense" suggests that the variability among plots differs among locations. Some locations may simply be more heterogeneous than others. In addition, plot size and experiment design may differ among locations. This understanding of the study sites argues for requiring different error variances for each location (Peterson, 1994).

Using environment-specific error variances complicates the analysis (Piepho, 2009). Environment-specific variances are estimated by restricted maximum likelihood (REML). The REML algorithm may fail because of infinite likelihood or failure to converge within the allowed number of iterations. The null distributions of *F* and *t* statistics have to be approximated by Satterthwaite or Kenward–Roger approximations. These may not maintain appropriate Type 1 error rates for hypothesis tests or appropriate coverage of confidence intervals (Richter et al., 2015).

When error variances differ among locations, the same common sense argument suggests that block variances should also be allowed to vary with location. If the small-scale heterogeneity among plots is not the same at two locations, then the medium-scale heterogeneity is also likely to differ among locations. Hence, the variance between blocks should be allowed to differ among locations. In our experience, this is rarely suggested and rarely done. SAS code to fit environment-specific block variances is in the supplemental material. Using environment-specific block variances in addition to environment-specific error variances further complicates the analysis.

The data can be used to assess equality of variances, but this has to be done carefully. Choosing between a pooled variance or separate variances for each environment is a specific example of the general problem of choosing a model for the variance-covariance matrix of the observations. This is commonly done using the Akaike Information Criterion, AIC, or related statistics based on the REML log-likelihood (Gbur et al. (2012), p. 80). The AIC statistic is computed for the model with a single pooled variance and compared with the AIC statistic for the model with separate variances for each environment. The model with the smaller AIC statistic is the more appropriate model. Models that have AIC statistics within 2 units of the smallest AIC are considered roughly comparable to the best model; models with AIC statistics that are more than 10 units larger than the best model are considered unlikely (Burnham and Anderson, 2002). The small-sample corrected AIC statistic (AICc), Schwartz Bayesian Criterion or other forms of Bayesian Information

Criterion (BIC) are sometimes used instead of AIC. Compared to AIC, BIC puts a larger penalty on the more complex model, so BIC more frequently selects simpler models, that is, the pooled variance model.

The operating characteristics of an AIC-based decision about pooling variances have not been well studied, but there is a strong argument that the use of AIC could be unduly sensitive to non-normality. AIC is similar to the likelihood ratio (LR) test and Bartlett's test of equal variances in that all three approaches depend on a log-likelihood that assumes normally distributed errors. Both the LR and Bartlett's tests are known to be very sensitive to violations of normality (Box 1953; Madansky 1988) so it seems likely that the AIC statistics will be also. An alternative is to use a test of equal variance that is less sensitive to normality, such as Levene's test (Madansky 1988, p. 67) or one of the many variations of Levene's test.

For the oat cultivar trial data, we consider four models for the error variance: one with a separate variance for each of the six environments (combinations of year and location), one with a separate variance for each year (but the same variance for all locations that year), one with a separate variance for each location (but the same variance in both years), and a pooled model with one error variance for all observations. The AIC statistics indicates that separate variances for each location (AIC = 698.4) should be used (Table 8). The AIC statistic for the pooled variance model (AIC = 701.1) is more than two units larger than the best model. The estimated location-specific variances are somewhat different (Ames: 17.6, Kawawha: 8.12, and Washington: 9.04), but they are not more than a factor of four different, which some use as a guide to assess when heterogeneity could be important (Fox 2008, p. 277). The Levene's test indicates no evidence of unequal variances between years, among locations, or in the year by location interaction (Table 8).

We suggest that in many cases, the use of a single pooled variance will be appropriate even when locations have somewhat different error variances unless the environment-specific sample sizes are very different. Our argument has three parts:

1. The overall treatment difference is an average of the environment-specific treatment differences. When variances are pooled, this is compared to the average error variance. This averaging leads to appropriate inferences, for the same reasons that an overall *F* statistic in a single-environment study is robust to variances that differ among treatments. Analogously, when sample sizes are not similar in the various environments, or if comparisons are to be made among specific pairs of environments, using a pooled error variance may be inappropriate.

TABLE 8. AIC statistics and Levene's test results to evaluate whether error variances can be pooled. The pooled model has one error variance for all observations. The location, year and year*location models fit separate error variances for each location, each year, or each combination of year and location.

Model	Number of variances	AIC	Levene's test results	
			Source	*P*-value
location	3	698.4	Year	0.79
pooled	1	701.1	Location	0.18
year*location	6	701.3	Year*Location	0.33
year	2	703.1		

2. In broad-sense inference, the variance that matters is the variance of an environment-specific treatment difference. This has two components: the error variance and the interaction variance. Either or both may differ between environments. When a single interaction variance is included in the model, the variances that matter will be more similar to each other than are the error variances.
3. Under narrow sense inference, treatment means and differences of treatment means are equally weighted averages of the environment-specific values, so the choice of pooled or not has no effect on the estimates. Under broad sense inference, with random interactions, the estimated means and their differences may or may not be equally weighted averages. When the variances are pooled, environment-specific estimates are equally weighted. When variances are separated, more weight is given to the estimates from the environments with smaller error variances. This situation is analogous to the distinction between Type II and Type III estimates and tests in fixed effects linear models. The choice of pooled or separate can change the estimates, but the desirable choice may not be clear.

To illustrate these points, we consider the three possible types of results in the oat cultivar trial: the overall mean harvest index for one cultivar, the overall mean difference between two cultivars, and the mean difference in one environment. We consider pooling error variances across all environments or using location-specific error variances with narrow-sense inference (fixed interaction, broad-sense inference (random environment by treatment interaction) and broad-sense inference with subdivided interactions (location by treatment, year by treatment, and location by year by treatment). SAS code to fit these models is in Box 6. The broad-sense model is fit without the nobound option because of convergence issues with the nobound model and default starting values.

Table 9 gives the estimates and standard errors when separate error variances are estimated for each location. The values in Table 9 can be compared to those using the pooled error variance (Table 6). When the focus is on the difference in the treatment means, the choice of pooled or separate error variances has almost no effect on the estimates or their standard errors (Table 9). One source of the small difference in standard errors is a different estimate of the treatment × environment variance component. When error variances are pooled, the interaction variance is estimated as 4.99; when error variances are estimated separately for each environment, the interaction variance is estimated as 5.47. When the focus is an environment-specific result, the choice of pooled or separate error variances affects both the estimates and their standard errors (Table 9), because of the much larger estimated error variance for the Ames location. If the sample sizes in each environment were substantially different, the choice of pooled or separate error variances would have a larger effect on the results.

Finally, we suggest that more attention needs to be given to treatment-specific error variances and treatment-specific components of the environment by treatment interaction. When the focus of a study is differences between treatments, the assumption that each treatment has the same error variance is much more important than the assumption that each environment has the same variance (Box, 1954).

BOX 6.

```
proc mixed nobound data=oats;
  class Environment Location Block Cultivar;
  model HI = Environment Cultivar Cultivar*Environment /ddfm=kr;
  random Block(Environment);
  repeated /group=Location;
  estimate ‘Cultivar: Cherokee’ intercept 1 Cultivar 1;
  estimate ‘Cultivar: Cherokee - Don’ Cultivar 1 -1;
  estimate ‘Cult: Cherokee - Don in Ames 1985’ Cultivar 1 -1
Environment*Cultivar 1 -1;
  title ‘Narrow sense inference, 3 location specific error
variances’;
  run;

proc mixed data=oats;
  class Environment Location Block Cultivar;
  model HI = Cultivar /ddfm=kr;
  random Environment Block(Environment) Cultivar*Environment;
  repeated /group=Location;
  estimate ‘Cultivar: Cherokee’ intercept 1 Cultivar 1;
  estimate ‘Cultivar: Cherokee - Don’ Cultivar 1 -1;
  estimate ‘Cultivar: Cherokee - Don in Ames, 1985’ Cultivar 1 -1
        | Cultivar*Environment 1 -1 ;
  title ‘Broad sense inference, 3 location specific error
variances’;
  run;
```

Some statisticians may take exception to our suggestions. We agree that ignorance of alternatives is a poor reason to pool. However, we encourage careful consideration of all sources of heteroscedasticity and return to Box’s (1954) conclusion for treatment comparisons, which is that treatment-specific variances are a much larger concern than environment-specific variances.

Choosing the Model for a Quantitative Treatment

When a treatment is a quantitative independent variable, such as rate of fertilizer or Julian day of planting, it can be modeled in two different ways. One is to treat each level of the quantitative variable as a separate group and use the approaches discussed previously. The other is to use a regression model to estimate a response curve (e.g., Cox, 1958, p. 30). When a simple regression model is an adequate fit to the data, regression provides especially straight-forward conclusions. A regression model also allows asking new questions about the effects of the quantitative treatment. Using a regression model in the analysis of a repeated experiment introduces new issues, which we now discuss.

To illustrate the issues, we consider a study of corn yield response to fertilizer addition on Antigua Island in the Lesser Antilles. The data are given in Table 58.1 in Andrews and Herzberg (1985). The 12 treatments were combinations of amounts of N, P, and K fertilizer using a 2 × 2 × 2 factorial treatment design supplemented by four additional treatments (Springer, 1972). Treatments were randomized to plots in an incomplete block design. The response variables are the number of ears and the total grain weight, in kg per plot. The entire design was repeated at eight locations. We focus on the response of grain weight to nitrogen (N) addition.

We use a subset of the data to illustrate the combined analysis of a regression model. We deleted all data from all plots at two locations (NSAN and WLAN) and certain plots at two additional locations because of experimental irregularities, detailed in Andrews and Herzberg (1985). A thorough evaluation of fertilizer effects would model the response to all three fertilizers, N, P, and K. We focus on the effects of N addition and simplify the example by omitting consideration of P and K. We remove all data from the treatments labeled 131 and 113, which have a high amount of P and high amount of K, respectively. After removing these two treatments, it is appropriate to model the effect of N addition independently of the effects of P or K.

A plot of the grain weight and fitted quadratic response curves at each location is in Fig. 3. Two locations, DBAN and LFAN, visually show a large effect of increasing N amount, two locations, OVAN, and ORAN, show a smaller effect, and two locations, WEAN and TEAN, show a very small effect. Heterogeneity in error variances between locations or between N levels was assessed by fitting a model with locations, blocks within location, N level as a factor, and N × location interactions, calculating the residuals from that fitted model, then using Levene's test on those residuals. There was no evidence of unequal variance between locations ($p = 0.46$) or between N levels ($p = 0.77$).

In a single experiment with replicated levels of a quantitative treatment, e.g., of N, the effect of that quantitative treatment can be evaluated using (at least) three different approaches. One is to fit a model with a separate mean for each level of N, and follow that up with multiple comparisons among treatment means. This approach ignores the quantitative nature of the treatment. A second approach is to fit a means model, but follow that up by estimating linear, quadratic, or more complicated contrasts (e.g., Gbur et al., 2012, p. 131–132; Vargas et al., 2018; Chapter 7) to summarize responses to N level. A third approach is to fit a regression model with N level as a continuous covariate. This model may be simple, e.g., a linear response to N, more

TABLE 9. Estimates and standard errors for the mean of one variety (Cherokee), a difference of two variety means (Cherokee – Don), and an environment-specific difference of two variety means (Cherokee – Don in Ames, 1995) when error variances are pooled across environments and when separate error variances are estimated for each location (three estimated error variances).

Quantity	Model	Estimate when error variances:		Standard error when error variances:	
		Pooled	Separate	Pooled	Separate
Mean:	Narrow sense	40.89	40.89	0.792	0.792
	Broad sense	40.89	40.81	1.721	1.734
	Broad subdivided	40.89	40.81	2.130	1.734
Difference:	Narrow sense	-9.50	-9.50	1.135	1.128
	Broad sense	-9.50	-9.41	1.718	1.745
	Broad subdivided	-9.50	-9.41	1.604	1.745
Difference in one environment	Narrow sense	-11.00	-11.00	2.780	3.370
	Broad sense	-10.35	-10.21	2.227	2.576
	Broad, subdivided	-10.45	-10.21	2.225	2.576

BOX 7.

```
proc import datafile = ‘antigua.csv’ out=antigua replace;
run;

proc mixed data=antigua;
  where location=’DBAN’;
  class N block;
  model wt = block N;
  lsmeans N / diff;
  title “Antigua, DBAN location, analysis of means for each N
level”;
  run;

proc mixed data=antigua;
  where location=’DBAN’;
  class N block;
  model wt = block N;
  estimate ‘linear N slope’ N -3 -1 1 3 /divisor=10;
  estimate ‘quadratic N slope’ N 1 -1 -1 1 /divisor=4;
  estimate ‘cubic N slope’ N -1 3 -3 1 / divisor=6;
  contrast ‘lack of fit to linear’ N 1 -1 -1 1, N -1 3 -3 1;
  title “Antigua, DBAN location, orthogonal polynomial trends”;
  run;

proc mixed data=antigua;
  where location=’DBAN’;
  class block;
  model wt = block N N*N;
  estimate ‘linear N slope’ N 1;
  estimate ‘quadratic N slope’ N*N 1;
  title “Antigua, DBAN location, regression analysis”;
  run;
```

complicated, e.g., a polynomial or nonlinear response to N, or nonlinear (discussed by Miguez et al., (2018) in Chapter 15). The regression model has the advantages of providing predictions of the response at N levels not used in the study and often being a more parsimonious description of the relationship between N level and the response. Box 7 shows SAS code to fit a means model, a means model with contrasts, and a polynomial regression to data from the DBAN location.

In a single experiment, these approaches differ in what is being described (means for each level or the overall pattern of response), how that pattern is estimated, and the definition of the error variance (pure error, i.e., only variability within replicate units at each level, or a combination of pure error and lack of fit to the proposed regression model). For example, when using linear or quadratic contrasts with a means model, the error variance is the pure error, but when using a regression model, the error variance is the combination of pure error and lack of fit. Contrasts and linear regression weight observations differently unless sample sizes are the same for each group. Linear regression weights each observation equally, so groups with large sample sizes get more weight. Contrasts weigh each mean equally, no matter what the sample size. Unless sample sizes are very unequal or the regression model has severe lack of fit, results from the means model followed by contrasts are similar to those from the regression model. We use the regression approach for the Antigua fertilizer response data.

For a repeated experiment, results from a means model and a regression model can be quite different because the two models describe experiment by treatment interactions differently, which leads to different variance–covariance matrices for the observations. For simplicity of exposition, we will focus on a linear response to N level and temporarily ignore all other features of the Antigua corn yield study. The means model for a repeated experiment is Eq. [3] with locations instead of days; the regression model for a repeated experiment is Eq. [5] in which location-specific linear regression models are jointly estimated.

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) X_{ij} + \varepsilon_{ij} \qquad [5]$$

where i indexes the location, j indexes the observation, and X_{ij} is the N level for observation j in experiment i. The b_{0i} terms quantify the between-location variability in the intercept of the regression. These are the equivalent of the day, that is, d_i, terms in Eq. [3]. The

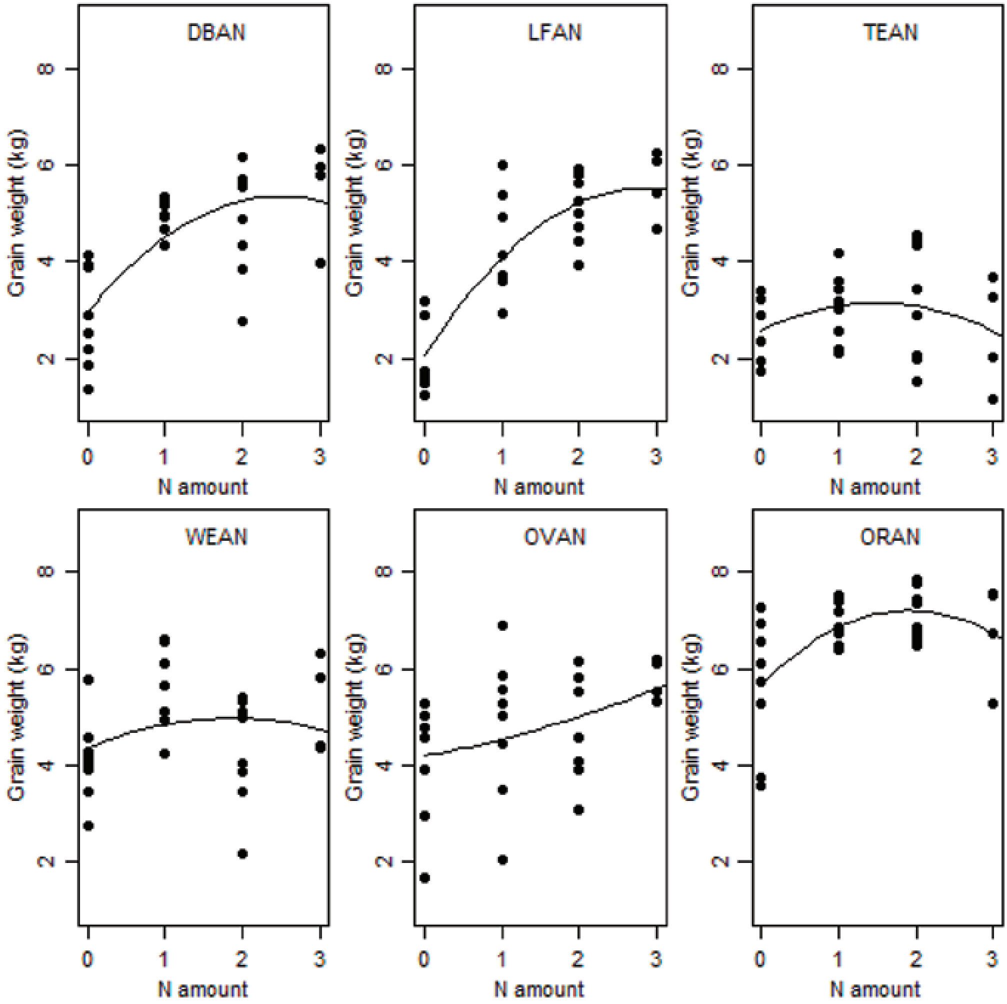

FIG. 3. Plot of grain weight vs N addition at the six Antigua locations used in the data analysis. Locations are indicated by their four letter abbreviations. The curve is the fitted quadratic from a separate analysis for each environment.

b_{1i} terms quantify the variability in the slopes of each regression; they are the equivalent of the day-by-treatment interaction terms in model (3). If every b_{1i} term is 0 in Eq. [5] the location-specific regression lines are parallel because they all have the same slope, β_1.

When broad-sense inference is used with Eq. [5], the location-specific intercepts and slopes become correlated random variables. These are commonly considered to be independent of the observation-specific errors. Hence, Eq. [5] becomes Eq. [6]:

$$Y_{ij}=(\beta_0+b_{0i})+(\beta_1+b_{1i})X_{ij}+\varepsilon_{ij} \qquad [6]$$

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}\right)$$

where σ_0^2 is the between-location variance in the intercepts, σ_1^2 is the between-location variance in the slopes, and is the covariance between the intercept and slope. Fitting model (Eq. [6]) is often numerically more stable after centering the X values by subtracting the overall mean X from all values. Centering the X values changes the interpretation of the intercept but does not change the interpretation of the slope.

The means model (Eq. [3]) and the regression model (Eqs. [5-6]) both allow location-to-location variation in the difference between two treatments, but they do so in different ways. The location-by-treatment interaction coefficients in Eq. [3] allow very general location-specific variations in the differences between two treatments. A non-zero interaction allows any sort of location variation. Two of the possible patterns with a non-zero interaction in Eq. [3] are shown in panels a and b of Fig. 4. The location-specific slope in Eq. [5] allows treatment differences to vary between locations but only in a very specific way. Consider the interaction between Treatments A and B in Locations 1 and 2. Under Eq. [3], that interaction is $(\mathrm{dt}_{1A}-\mathrm{dt}_{1B})-(\mathrm{dt}_{2A}-\mathrm{dt}_{2B})$. According to Eq. [5], that difference is $(\beta_1+b_{11})(X_A-X_B)-\left[(\beta_1+b_{12})(X_A-X_B)\right]=(b_{11}-b_{12})(X_A-X_B)$. The interaction depends on the difference in the X values and the difference in the two location-specific slopes. Two of the possible patterns with a non-zero interaction in Eq. [5] are shown in panels c and d of Fig. 4.

How should you determine which is the more appropriate model for the data? The commonly-used data-based approach is to compare AIC statistics. Because the means model 3, (Eq. [3]) and the regression model 6 (Eq. [6]) have different fixed effects, the AIC statistics must be computed using ML, not the default REML. The other approach is based on your understanding of how the experiments might differ from each other. If the relationship between the response and an X variable is based on a well-understood mechanism and it is reasonable that different experiments will have different regression slopes, then the random coefficients regression, model 6 (Eq. [6]), is the more appropriate model. If the relationship between the response and an X variable is expected to be more complex in each experiment, then perhaps the general experiment-by-treatment interaction, model 3 (Eq. [3]) is more appropriate.

A regression model for repeated experiments requires both specifying the form of the regression function, for example, linear or quadratic, and choosing which coefficients vary across the experiments. The need for location-specific linear or quadratic slopes can be evaluated using either the narrow-sense or broad-sense model. We find it easier to use a narrow-sense model for model selection. The form of the N response was evaluated by the SAS code given in Box 8. The *p*-value for the quadratic effect of N is very small, but the *p*-value for the quadratic by location interaction is 0.48, so that interaction will be dropped. When the narrow-sense model is re-run without the quadratic by location interaction, the *p*-values for the quadratic N effect and the linear by location interaction are 0.0010 and 0.0001, respectively, so no further simplifications are made to the model. The location-specific plots of residuals against predicted values indicate no concerns.

Code to fit the broad-sense model is in Box 9. The estimated variance components are 1.62 for the locations, 0.07 for blocks within locations, 0.16 for the linear nitrogen slopes, and 0.96 for the pooled error variance. The estimated location-specific intercepts and

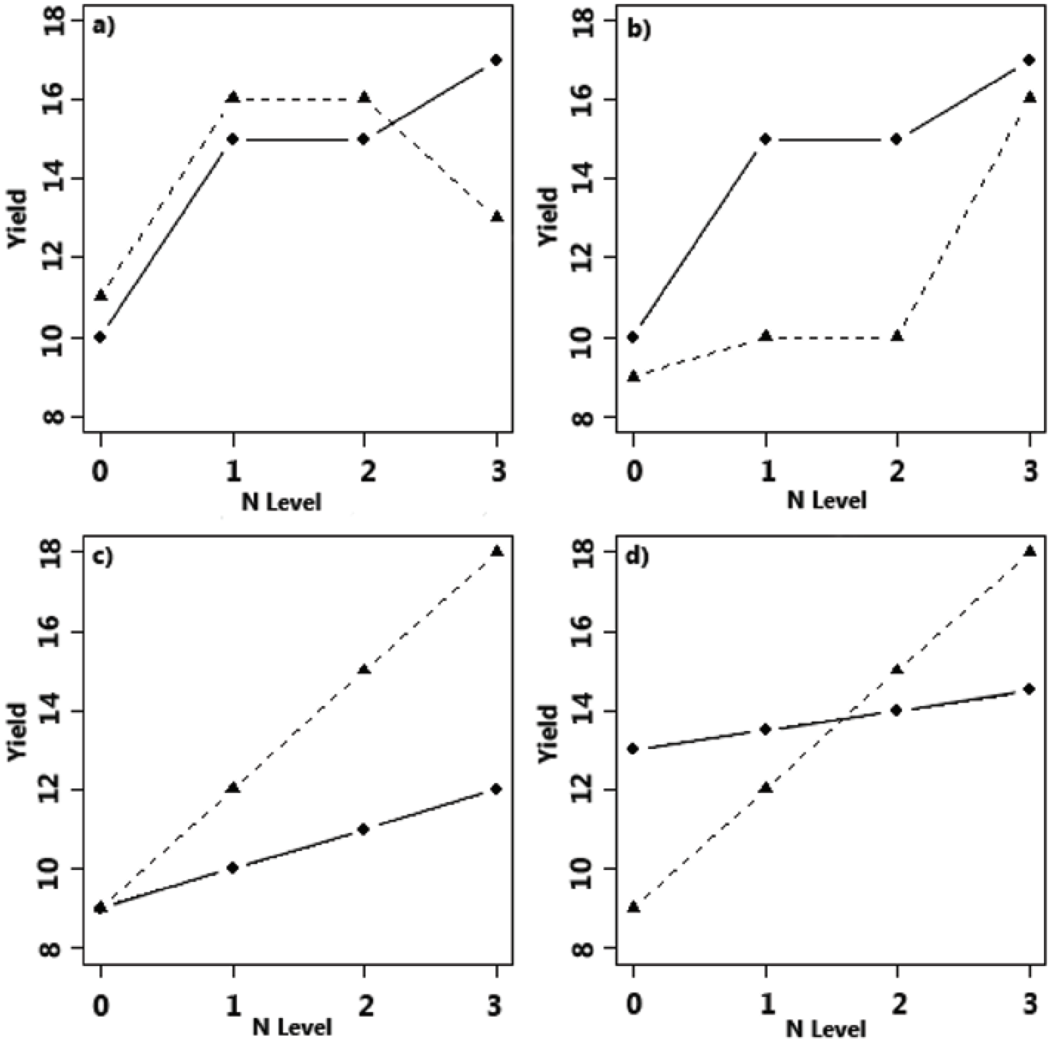

FIG. 4. Location by treatment Interaction patterns allowed by the means model (panels a and b) and the linear regression model (panels c and d). Treatment means for Location 1 are shown with circles and solid lines; treatment means for Location 2 are shown with triangles and dashed lines. Interaction in the means model (equation 3) allows any pattern of non-parallel responses to treatment. Interaction in the regression model (Eqs. 5 or 6) only allows patterns that correspond to different regression slopes.

BOX 8.

```
proc mixed data=antigua;
  class location block;
  model wt = N N*N N*location N*N*location / ddfm=kr;
  random location block(location);
  title ‘Quantitative N, with quadratic interaction’;
  run;

proc mixed data=antigua;
  class location block;
  model wt = N N*N N*location / ddfm=kr;
  random location block(location);
  title ‘Quantitative N, only linear N*location’;
  run;
```

BOX 9.

```
/* preceding code, not in box 9, creates prediction locations */
proc mixed data=full;
  class location block;
  model wt = N N*N / outpm=predmean outp=predloc ddfm=kr;
  random intercept N /subject=location type=un;
  random block(location);

  estimate ‘intercept’ intercept 1;
  estimate ‘N linear’ N 1;
  estimate ‘N quadratic’ N*N 1;

  title ‘Quantitative treatments’;
  run;
```

location-specific slopes are correlated. The estimated covariance is -0.18, so the estimated correlation is -0.36. Estimates and tests of the fixed effects describing the average response curve are shown in Table 10. More important are the predictions of the location-specific N response curves shown in Table 11. We estimate the parameters of the overall curve, because those are fixed effects. We predict the coefficients for each location because the location-specific coefficients are random variables. Because the model did not include a location by quadratic interaction, the quadratic coefficient is the same for all six locations. The predicted intercept and slope do vary between the locations.

The overall N response curve and the location-specific response curves are plotted in Fig. 5. The standard errors for the overall curve are larger than those for a location-specific curve because the uncertainty in the overall curve includes the variation among the six locations.

Summary

Combined analyses of repeated experiments provide estimates of both average treatment effects and their consistency across environments. Doing a combined analysis requires decisions about the scope of inference, pooling of error variances across environments, and pooling of contributions to treatment by environment interactions. The scope of inference (narrow- or broad-sense) has major consequences for the analysis. Narrow-sense conclusions apply to the set of environments in the study; broad-sense conclusions apply to a more general population of environments. Error variances are likely to differ across environments, especially when environments are different locations. Conclusions about environment-specific treatment effects depend on the choice of pooled or environment-specific error variances, but conclusions about average treatment effects are quite robust to that choice. The choice of pooling treatment by environment interactions is especially important when treatments differ in their consistency across environments. The analysis may also require deciding whether to model quantitative treatments using means models or regression models.

TABLE 10. Estimates and tests for fixed effects in the broad-sense model for the Antigua response to fertilization study. Denominator df are computed using the Kenward–Rogers approximation.

Parameter	Estimate (se)	Denominator df	*P*-value
Intercept	3.64 (0.54)	5.17	0.001
N	1.28 (0.29)	28.6	< 0.0001
N*N	-0.267 (0.079)	135	0.001

TABLE 11. Predictions of the location-specific coefficients for N response.

Location	Intercept	N	N*N
DBAN	3.14	1.54	-0.267
LFAN	2.39	1.86	-0.267
ORAN	5.70	1.21	-0.267
OVAN	3.90	1.20	-0.267
TEAN	2.53	0.87	-0.267
WEAN	4.19	0.98	-0.267

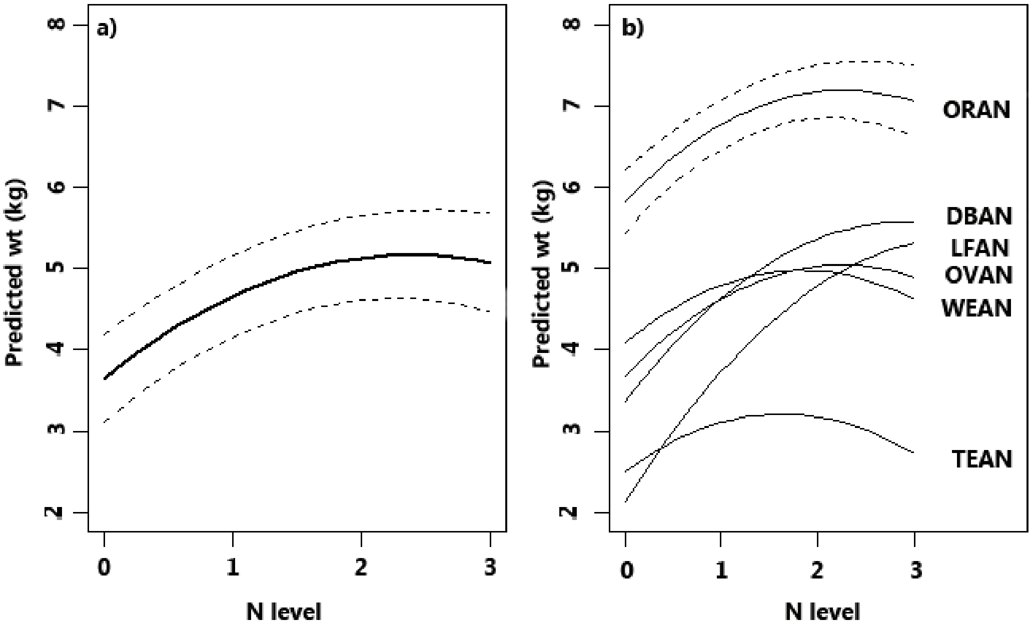

FIG. 5. Overall mean N response curve (panel a) and location-specific predicted N response curves (panel b). Dashed lines indicate +/- 1 standard error around the overall curve and the location-specific curve for the ORAN location.

Key Learning Points:

- What is the difference between broad- and narrow-sense inference about treatment effects?
- How do you specify a model to obtain narrow-sense inference? Broad-sense inference?
- What are consequences, benefits, and disadvantages of using environment-specific variances?
- How do you specify a model with environment-specific error variances?
- How can you decide whether to pool or subdivide treatment by environment interactions when doing broad-sense inference?
- How can a regression model be extended to multiple environments?

Review Questions (T/F)

For all questions, consider an experiment comparing three tillage treatments repeated in 18 environments. The study area at each location is divided into three blocks with three plots in each block. Tillage treatment is randomly assigned to plot within each block. Different plots are used in each environment.

1. In this study, modeling environment effects and block effects as fixed effects or as random effects leads to the same inference about the difference between the two types of tillage.
2. If conclusions about the effects of tillage in a new location are desired, you should use narrow sense inference.
3. Narrow sense conclusions about the effects of tillage usually have smaller standard errors than do broad sense conclusions.
4. To obtain narrow-sense conclusions, omit the treatment by environment interaction from the model.
5. To obtain broad-sense conclusions, define the treatment by environment interaction as a random effect.
6. Broad-sense confidence intervals for the difference between two types of tillage will be based on *t* distributions with 18 degrees of freedom.
7. The combined analysis across environments requires that plot-plot variation be pooled across environments.
8. The combined analysis across environments requires that variation between blocks be pooled across environments.
9. The 18 environments are actually six locations, each studied for three years. Tillage effects are expected to vary somewhat among locations because of different soil characteristics. Tillage effects are not expected to vary among years at the same location. In this case, subdividing the treatment by environment interaction will have minimal effect on the conclusions about the treatment effect.
10. Imagine that the three tillage treatments are three levels of some quantitative factor, for example, amount of soil disturbance. The data for each environment could be

analyzed using a regression model with a linear effect of soil disturbance. It is possible to construct a combined analysis of those regression models in all 18 environments.

Exercises

1. An experiment was conducted to assess the effect of a fungicide treatment on soybean yield (kg ha^{-1}). It was conducted as an on-farm strip-plot trial with six pairs of side-by-side strips of which one randomly received fungicide treatment. The experiment was repeated at eleven farms (environments). The data were extracted from a much larger dataset provided by the Iowa Soybean Association and are provided in the on-farm soybean dataset in the supplemental materials.

   a. Analyze the experiment separately for each environment.

   b. Evaluate the error variances to determine whether or not they may be considered homogeneous.

   c. You are interested in broad-sense inferences about the average difference between the fungicide and control treatments. Conduct a combined analysis assuming environment and replication to be random factors and treatment as fixed.

   d. Interpret the results of the experiment with respect to the efficacy of fungicide treatment in improving soybean yield.

2. Antonio Mallorino at Iowa State University has studied corn response to P fertilization since 2002. The Prate.csv file contains 13 yr of data from the SouthWest research farm. The design is a RCBD with three blocks of five plots each. Four P levels (0, 28, 56, and 112 lb ac^{-1}) were used; the 0 level was replicated twice in each block. Blocks and plots can be considered independent across years. The response variable is yield in bu acre^{-1}.

Consider years to be a fixed factor and Prate to be a continuous variable.

   a. What sort of polynomial model is appropriate to describe yield response to Prate? Linear? Quadratic? With one coefficient for all years, or coefficients that differ among years?

Now consider years to be a random factor. Fit a quadratic model that allows the intercept and linear Prate coefficient to vary between years (but the quadratic coefficient is constant).

   b. What is the equation that predicts yield as a function of Prate for a year not in the data set, (e.g., 2015)?

   c. What is the year-to-year variability in the linear Prate slope? Use the standard deviation to describe that variability.

   d. Examine the residuals. Is it appropriate to use yield as the response variable, or should yield be transformed?

e. Apply Levene's test to the residuals to assess whether the error variance differs among years.

f. Refit the model used in Parts b through e with year-specific error variances. Do the answers to Parts B and C change much?

References

Andrews, D.F., and A.M. Herzberg. 1985. Data: A collection of problems from many fields for the student and research worker. Springer, New York. doi:10.1007/978-1-4612-5098-2

Bennett, C.A., and N.L. Franklin. 1954. Statistical analysis in chemistry and the chemical industry. John Wiley & Sons, New York.

Binns, M.R., P.M. Morse, and B.K. Thompson. 1983. Re: "Analysis of combined experiments" by M.S. McIntosh. Agron. J. 75:1056. doi:10.2134/agronj1983.00021962007500060045x

Blouin, D.C., E.P. Webster, and J.A. Bond. 2011. On the analysis of combined experiments. Weed Technol. 25(01):165–169. doi:10.1614/WT-D-10-00047.1

Box, G.E.P. 1953. Non-normality and tests on variances. Biometrika 40:318–335. doi:10.1093/biomet/40.3-4.318

Box, G.E.P. 1954. Some theorems on quadratic forms applied to the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. Ann. Math. Stat. 25:484–498. doi:10.1214/aoms/1177728717

Burnham, K.P., and D.R. Anderson. 2002. Model selection and multimodel inference, 2nd ed., Springer, New York.

Carmer, S.G., W.M. Walker, and R.D. Seif. 1969. Practical suggestions on pooling variances for F tests of treatment effects. Agron. J. 61(2):334–336. doi:10.2134/agronj1969.00021962006100020051x

Casler, M.D. 2018. Blocking principles for biological experiments. In: B. Glaz and K.M. Yeater, Applied statistics in agricultural, biological,and environmental sciences. ASA, CSSA, SSSA, Madison, WI.

Cochran, W.G., and G.M. Cox. 1957. Experimental designs, 2nd ed. John Wiley & Sons, New York.

Cox, D.R. 1958. Planning of experiments. John Wiley & Sons, New York.

Diggle, P.J., P. Heagerty, K.-Y. Liang, and S.L. Zeger. 2002. Analysis of longitudinal data, 2nd ed. Oxford University Press, Oxford, U.K.

Dixon, P. 2016. Should blocks be fixed or random? Annual Conference on Applied Statistics in Agriculture, Manhattan, KS. 1-3 May 2016. New Prairie Press, Manhattan, KS. http://newprairiepress.org/agstatconference/2016/proceedings/4 (verified 17 Nov. 2017).

Edwards, A.L. 1985. Experimental design in psychological research, 5th ed., Harper and Row, New York.

Fisher, R.A. 1935. The design of experiments. Oliver and Boyd, Edinburgh.

Fox, J. 2008. Applied regression analysis and generalized linear models, 2nd ed. Sage, Thousand Oaks, CA.

Gauch, H.G. 2006. Statistical analysis of yield trials by AMMI and GGE. Crop Sci. 46:1488–1500. doi:10.2135/cropsci2005.07-0193

Gbur, E.E., W.W. Stroup, K.S. McCarter, S. Durham, L.J. Young, M. Christman, M. West, and M. Kramer. 2012. Analysis of generalized linear mixed models in the agricultural and natural resources sciences. Agronomy Society of America, Madison, WI.

Gelman, A. 2005. Analysis of variance: Why it is more important than ever. Ann. Stat. 33(1):1–53. doi:10.1214/009053604000001048

Giesbrecht, F.G. 1989. A general structure for the class of mixed linear models. Application of mixed models in agriculture and related disciplines. Louisiana Agricultural Experiment Station, Baton Rouge, LA. p. 183-201.

Giesbrecht, F.G., and J.C. Burns. 1985. Two-stage analysis based on a mixed model: Large-sample asymptotic theory and small-sample simulation results. Biometrics 41:477–486. doi:10.2307/2530872

Giesbrecht, F.G., and M.L. Gumpertz. 2004. Planning, construction, and statistical analysis of comparative experiments. John Wiley & Sons, New York. doi:10.1002/0471476471

Hocking, R.R. 1985. The analysis of linear models. Brooks/Cole, Monterey, CA.

Kenward, M.G., and J.H. Roger. 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 53:983–997. doi:10.2307/2533558

Keppel, G., and T.D. Wickens. 2004. Design and analysis: A researcher's handbook. Pearson, Prentice Hall, Upper Saddle River, NJ.

Koricheva, J., J. Gurevitch, and K. Mengersen, editors. 2013. Handbook of meta-analysis in ecology and evolution. Princeton Univ. Press, Princeton NJ. doi:10.1515/9781400846184

Laird, R.J., and F.B. Cady. 1969. Combined analysis of yield data from fertilizer experiments. Agron. J. 61:829–834. doi:10.2134/agronj1969.00021962006100060001x

Madansky, A. 1988. Prescriptions for working statisticians. Springer-Verlag, New York. doi:10.1007/978-1-4612-3794-5

McCulloch, C.E., S.R. Searle, and J.M. Neuhaus. 2008. Generalized, linear, and mixed models. 2nd ed. John Wiley & Sons, New York.

McIntosh, M.S. 1983. Analysis of combined experiments. Agron. J. 75:153–155. doi:10.2134/agronj1983.00021962007500010041x

McLean, R.A., W.L. Sanders, and W.W. Stroup. 1991. A unified approach to mixed linear models. Am. Stat. 45:54–64.

Miguez, F., S. Archontoulis, and H. Dokoohaki. 2018. Nonlinear regression models and applications. In: B. Glaz and K.M. Yeater, editors, Applied statistics in agricultural, biological, and environmental sciences. ASA, CSSA, SSSA, Madison, WI.

Moore, K.J., and P.M. Dixon. 2015. Analysis of combined experiments revisited. Agron. J. 107:763–771. doi:10.2134/agronj13.0485

Nelder, J.A. 1965a. The analysis of randomized experiments with orthogonal block structure. I Block structure and the null analysis of variance. Proc. R. Soc. London, Ser. A 283: 147-162.

Nelder, J.A. 1965b. The analysis of randomized experiments with orthogonal block structure. II Treatment structure and the general analysis of variance. Proc. R. Soc. London, Ser. A 283: 163-178.

Newman, J.A., J. Bergelson, and A. Grafen. 1997. Blocking factors and hypothesis tests in Ecology: Is your Statistics text wrong? Ecology 78:1312–1320. doi:10.1890/0012-9658(1997)078[1312:BFAHTI]2.0.CO;2

Patterson, H.D. and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545-554.

Pearce, S.C. 1988. Analysis of data from agricultural experiments. Trop. Agric. (St Augustine) 65(1):2–5.

Peterson, R.G. 1994. Agricultural field experiments: Design and analysis. Marcel Dekker, New York.

Piepho, H.-P. 2009. Data transformation in statistical analysis of field trials with changing treatment variance. Agron. J. 101:865–869. doi:10.2134/agronj2008.0226x

Piepho, H.P., A. Büchse, and K. Emrich. 2003. A hitchhiker's guide to mixed models for randomized experiments. J. Agron. Crop Sci. 189:310–322. 10.1046/j.1439-037X.2003.00049.x

Piepho, H.P., A. Büchse, and C. Richter. 2004. A mixed modelling approach for randomized experiments with repeated measures. J. Agron. Crop Sci. 190:230–247. doi:10.1111/j.1439-037X.2004.00097.x

Piepho, H.-P., J. Möhring, T. Schulz-Streeck, and J.O. Ogutu. 2012. A stage-wise approach for the analysis of multi-environment trials. Biom. J. 54:844–860. doi:10.1002/bimj.201100219

Piepho, H.P., E.R. Williams, and M. Fleck. 2006. A note on the analysis of designed experiments with complex treatment structure. HortScience 41:446–452.

Richter, C., B. Kroschewski, H.-P. Piepho, and J. Spilke. 2015. Treatment comparison in agricultural field trials accounting for spatial correlation. J. Agric. Sci. 153:1187–1207. doi:10.1017/S0021859614000823

Robinson, G.K. 1991. That BLUP is a good thing– the estimate of random effects. Stat. Sci. 6:15–32. doi:10.1214/ss/1177011926

Robinson, D.L., Thompson, R. & Digby, P.G.N. 1982. REML − a program for the analysis of non-orthogonal data by restricted maximum likelihood. In: H. Caussinus and P. Ettinger, Compstat 1982 Proceedings in Computational Statistics, Part II (supplement), 231-232. Physica-Verlag, Vienna, Austria.

Searle, S.R. 1971. Linear models. John Wiley & Sons, New York.

Springer, B.G.F. 1972. Experimental design and analysis under limited resources. Proceedings of the Caribbean Food Crops Society, Tenth annual meeting, San Juan, Puerto Rico. 16 June 1972. Caribbean Food Crops Society, Mayaguez, Puerto Rico. p. 147-151.

Stroup, W.W. 1989. Why mixed models? Application of mixed models in agriculture and related disciplines. Louisiana Agricultural Experiment Station, Baton Rouge, LA. p. 1-8.

Stroup, W.W. 2013. Generalized linear mixed models: Modern concepts, methods and applications. CRC Press, Boca Raton, FL.

Sulc, R.M., E. van Santen, K.D. Johnson, C.C. Sheaffer, D.J. Undersander, L.W. Bledsoe, D.B. Hogg, and H.R. Willson. 2001. Glandular-haired cultivars reduce potato leafhopper damage in alfalfa. Agron. J. 93:1287–1296. doi:10.2134/agronj2001.1287

Thompson, R.W., H.A. Fribourg, J.C. Waller, W.L. Sanders, J.H. Reynolds, J.M. Phillips, S.P. Schmidt, R.J. Crawford, V.G. Allen, and D.B. Faulkner. 1993. Combined analysis of tall fescue steer grazing studies in the Eastern United States. J. Anim. Sci. 71:1940–1946. doi:10.2527/1993.7171940x

Thompson, S.K. 1992. Sampling. John Wiley & Sons, New York.

Vargas, M., B. Glaz, J. Crossa, and A. Morgounov. 2018. Analysis and interpretation of interactions of fixed and random effects. In: B. Glaz and K.M. Yeater, editors, Applied statistics in agricultural, biological, and environmental sciences. ASA, CSSA, SSSA, Madison, WI.

Wilkinson, G.N. 1970. A general recursive algorithm for analysis of variance. Biometrika 57: 19-46.

Wright, K. 2013. Revisiting Immer's barley data. Am. Stat. 67:129–133. doi:10.1080/00031305.2013.801783

Yates, F. 1937. The design and analysis of factorial experiments of the Commonwealth Bureau of Soil Science. Technical Communication No. 35. Commonwealth Agricultural Bureaux, Farnham Royal, UK.