Journal of Agricultural Science (2006), **144**, 393–409. © 2006 *Cambridge University Press*
doi:10.1017/S0021859606006319 *Printed in the United Kingdom*

The design and analysis of multi-phase plant breeding experiments

A. B. SMITH[1]*, P. LIM[2] AND B. R. CULLIS[1]

[1] *Wagga Wagga Agricultural Institute, Private Mail Bag, Wagga Wagga, NSW, Australia 2650*
[2] *Charles Sturt University, Wagga Wagga*

(*Revised MS received 1 May 2006; First published online 4 September 2006*)

SUMMARY

Despite the importance of selection for quality characteristics in plant improvement programmes, literature on experimental design and statistical analysis for these traits is scarce. Most quality traits are obtained from multi-phase experiments in which plant varieties are first grown in a field trial then further processed in the laboratory. In the present paper a general mixed model approach for the analysis of multi-phase data is described, with particular emphasis on quality trait data that are often highly unbalanced and involve substantial sources of non-genetic variation and correlation. Also detailed is a new approach for experimental design that employs partial replication in all phases. The motivation for this was the high cost of obtaining quality trait data, thus the need to limit the total number of samples tested, but still allow use of the mixed model analysis. A simulation study is used to show that the combined use of the new designs and mixed model analysis has substantial benefits in terms of the genetic gain from selection.

INTRODUCTION

Breeding for improved quality is an important aspect of plant improvement programmes. It is therefore crucial to obtain accurate and reliable phenotypic information, both for the purposes of varietal selection and identification of quantitative trait loci (QTL). Experimental design and statistical analysis have a key role to play in this quest. The benefits of sound design and analysis are well documented in terms of the key trait of grain yield. There are numerous articles devoted to methods of field trial design and the statistical analysis of the resultant yield data. The adoption of efficient techniques is reasonably widespread and the role of replication and randomization and the notion of spatial heterogeneity are well understood and accepted. The scenario for quality traits is very different. Key quality traits, such as flour yield for wheat and malting quality for barley, are obtained from multi-phase experiments (see Brien 1983, for example). Specifically, grain samples are taken from a field experiment (Phase I) then processed further in a laboratory experiment (Phase II) or sequence of experiments (Phases II and higher). The level of adoption of sound design and analysis techniques in multi-phase plant breeding experiments is low. For example, the use of replication is the exception rather than the rule. Common practice involves the use of a single field replicate (or composite of several replicates) of varieties and no randomization or replication of grain samples in the laboratory (although a laboratory control sample is often processed at regular intervals). It is also common that raw (un-analysed) data is used for selection and QTL identification.

In the present paper, the design and analysis of multi-phase quality trait data are discussed in the context of variety trials conducted by plant breeding and evaluation programmes. The data may therefore relate either to early generation breeding trials or late stage evaluation trials. It is the present authors' opinion that in both cases the aim is to maximize the genetic gain from selection of a subset of superior varieties (also referred to as genotypes). In early generation trials, selection is undertaken by breeders in order to progress the best genotypes through the breeding programme. In final evaluation trials, selection takes the form of decisions by farmers and advisors regarding the best genotypes for commercial use. Thus the aim, and thence the approach to design and analysis, is the same for all stages of testing. The

* To whom all correspondence should be addressed. Email: alison.smith@dpi.nsw.gov.au

reader is referred to Smith *et al.* (2005) for a more comprehensive discussion of this issue.

The literature on the analysis of quality trait data from multi-phase plant breeding experiments is scarce. In a general setting, key references for multi-phase experiments are McIntyre (1955), Wood *et al.* (1988), Brien (1983) and Brien & Bailey (in press). Those authors use analysis of variance techniques that aim to appropriately account for the block structure in the experiment. As noted by Wood *et al.* (1988), 'The distinctive feature of two-phase [or multi-phase] experiments is that each phase has its own block structure and these must be combined to form the overall variance model.' In single phase experiments determination of the block structure is usually straightforward (see the seminal paper of Nelder 1965) but it may be more difficult in the context of multi-phase experiments. Brien (1983) provides some helpful guidelines with the concept of 'tiers'. The nature of quality trait data is such that there often exist sources of variation and correlation additional to that accounted for by the block structure. In the present paper, therefore, a mixed model analysis that incorporates a modelling aspect is proposed. The approach is based on that of Smith *et al.* (2001*a*) and Cullis *et al.* (2003) and has been motivated by experience in analysing two key quality traits, namely milling yield of wheat and malting quality of barley.

In terms of experimental design for multi-phase quality data, the focus in the present paper is on the second (and higher) phases. Thus, it is assumed that the field trial has already been conducted and, given the field layout, a design is required for the laboratory phase(s). It may seem more desirable to construct the designs for all phases simultaneously, as is done in some other areas of research that employ multi-phase testing (manufacturing and the food industry, for example). In these settings the number of treatments is relatively small and often comprise factorial combinations where different treatment factors are applied in different phases. The design for such experiments is well accepted and involves standard techniques. The application under consideration in the present paper is more complex. An important issue is that it is common for only a subset of the varieties grown in the field trial to be quality tested. Additionally, the subset is not known prior to the conduct of the field trial. Thus, it is impossible to design the field phase in conjunction with the laboratory phases.

With respect to the design for the field phase of plant breeding experiments, there are differences between the designs used for early and late stage testing. Historically, grid-plot designs (with single plots of test varieties and multiple plots of standard varieties arranged in a grid throughout the trial) have been used for early stage trials. More recently, Cullis *et al.* (in press) proposed an alternative that uses replication of a percentage p of test varieties as a replacement for the multiple grid plots. Cullis *et al.* (in press) showed that, for a fixed total number of field plots, their partially replicated designs resulted in higher genetic gains than grid-plot designs. In later stages of testing fully replicated designs are standard practice. In both the partially and fully replicated designs, the use of blocking and strategies to accommodate spatial trend are recommended.

In terms of the design for the laboratory phases, a key issue is the amount of replication. It can be shown (see Kempton 1984, for example) that for given levels of genetic and error variance and a fixed amount of resources, increasing the number of replicates (and therefore reducing the number of genotypes) will result in a reduction in expected genetic gain (due to the reduction in selection intensity). Thus, for this scenario, genetic gain is maximized with the use of an experimental 'design' with no replication, then basing selection on the raw data. The absence of replication, however, precludes a statistical analysis. If an appropriate analysis causes a reduction in effective error variance then an increase in genetic gain may result. Thus, knowledge and understanding of the sources of variation affecting quality traits is essential in order to make recommendations about experimental design. This can only be achieved by conducting appropriate statistical analyses of suitable data-sets, namely data characterized by replication and randomization in all phases. Such data are relatively rare. Since the work of Smith *et al.* (2001*a*) and Cullis *et al.* (2003), however, a substantial number of wheat milling experiments (and a lesser number of barley malting quality experiments) in Australian plant breeding and molecular marker programmes have incorporated some level of field and laboratory replication. This has enabled mixed model analyses of the data to be conducted. In the majority of cases the modelling of non-genetic (error) variation resulted in reductions in effective error variance. In the present paper, therefore, an approach for the design of multi-phase quality experiments that uses replication in all phases of the experiment is proposed, thereby allowing a rigorous statistical analysis of the data with the expectation of an associated increase in genetic gain.

The paper is arranged as follows. In the following section, 10 motivating multi-phase quality trait data-sets are described. In the section 'Statistical analysis', an approach for the statistical analysis of such data is presented, then applied to the motivating examples. The new designs are described in the section 'A new class of experimental design' and assessed via a simulation study. Concluding remarks are presented in the Discussion.

MOTIVATING EXAMPLES

Here, 10 wheat milling data-sets are considered. They are typical of Australian plant breeding milling data.

Table 1. *NSW field trials: numbers of genotypes and plots*

	Genotypes			Plots			
		Control			Control		Layout
Trial	Test	Grid	Other	Test	Grid	Other	Columns × Rows
NSW-1	426	6	4	426	111	15	12 × 46 = 552
NSW-2	443	6	4	443	113	8	12 × 47 = 564
NSW-3	1415	5	7	1415	371	38	12 × 152 = 1824
NSW-4	1464	5	7	1464	382	38	12 × 157 = 1884
NSW-5	794	5	10	794	204	22	12 × 85 = 1020

Each is characterized by having some degree of replication in both the field and milling phases of the experiment. This allows a statistical analysis to be conducted in which the major sources of variation and correlation affecting the trait can be accommodated. Five of the data-sets relate to early generation trials from the NSW Department of Primary Industries (NSWDPI) wheat breeding programme. The remaining five data-sets relate to mapping populations from the Australian Winter Cereals Molecular Marker Programme (AWCMMP). Unlike the NSW data, the aim of these experiments was not varietal selection but the detection of QTL. They are considered here, however, since the phenotypic data still require a statistical analysis and the methods proposed for selection trials are also appropriate in this setting (see Eckermann *et al.* 2001; Verbyla *et al.* 2003, for example).

In terms of the NSW data-sets, all field trials were grid plot designs in which there was a single plot for each test genotype and multiple plots of control varieties arranged in a systematic grid throughout the trial. Additionally, there were multiple plots of some other commercial control varieties. Each trial was laid out in a rectangular array indexed by field rows and columns. Details of the field trial layouts are given in Table 1.

The milling of grain samples from a field trial is conducted as a sequential process that usually requires more than a single day of processing. Each grain sample is ground in a mill until all bran has been removed and the endosperm has been reduced to flour. The data of interest here is the flour yield, which is the weight of flour expressed as a proportion of the weight of the original grain sample. Grain samples from the NSWDPI trials were milled in the Wagga Wagga Agricultural Institute cereal chemistry laboratory using a Quadramat Junior mill. The plots milled in the laboratory corresponded to the subset of test genotypes of interest to the breeder (so some genotypes had already been discarded on the basis of other traits such as grain yield) and a sample of control plots chosen on the basis of their spatial location to provide reasonable coverage of the field trial. A proportion of the plots was replicated in the laboratory. These usually corresponded to the test genotypes. In designing the milling trial, all samples were randomized to days and times of processing within days. This involved a two-stage process. First, a design was generated for the replicated samples only, involving a reduced layout compared with the full set of samples (same number of days but fewer samples per day as appropriate). The design for these samples was a resolvable incomplete block design with milling days being used as the block factor. A complete replicate was milled in the first half of the trial (i.e. on days 1 to $d/2$, where d is the total number of days for the trial); the other replicate in the second half. This design was then expanded to the full dimensions by increasing the number of samples per day (achieved by inserting samples between those occupied by replicated samples) and allocating the remaining samples to the vacant positions at random. Details of the milling trial layouts are given in Table 2. Note that the proportion of each field trial milled in the laboratory ranged from 0·52 of the plots for NSW-5 down to 0·29 for NSW-3. The number of control plots milled as a proportion of the total number of plots milled was fairly consistent, ranging from 0·15 for NSW-3 down to 0·10 for NSW-4. This represents the field replication present in the milling process. In terms of laboratory replication, the proportion of field plots that were tested with laboratory replication ranged from 0·32 for NSW-5 down to 0·20 for NSW-1.

In terms of the AWCMMP data-sets, all field trials were designed as RCB with either two or three replicates (with additional plots of control varieties in some cases). The genotypes comprised doubled haploid (DH) genotypes together with their parental varieties and sometimes commercial control varieties. Details of the field trial layouts are given in Table 3.

Grain samples from the first four DH trials were milled using a Buhler mill and the last using a Quadramat Junior mill. In contrast to the NSW data, all (or nearly all) field plots were milled for all trials

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

Table 2. *NSW milling trials: numbers of genotypes and plots (milled either as single samples or replicated)*

Trial	Genotypes		Test plots		Control plots		Total plots	Layout
	Test	Control	Singles	Rep'd	Singles	Rep'd		Days × Times
NSW-1	214	7	167	47	33	3	250	10 × 30 = 300
NSW-2	201	7	136	65	34	0	235	10 × 30 = 300
NSW-3	449	7	340	109	78	1	528	22 × 29 = 638
NSW-4	622	7	412	210	68	0	690	30 × 30 = 900
NSW-5	463	7	295	168	69	0	532	14 × 50 = 700

Table 3. *DH field trials: numbers of genotypes and plots*

Trial	Genotypes		Plots		Layout
	DH	Other	DH	Other	Columns × Rows
DH-1	144	4	288	8	8 × 37 = 296
DH-2	174	8	348	24	12 × 31 = 372
DH-3	181	2	543	33	12 × 48 = 576
DH-4	175	8	350	22	12 × 31 = 372
DH-5	116	4	348	12	12 × 30 = 360

except DH-3, in which two out of the three field plots for each genotype were milled (see Table 4). In the first three trials, none of the field plots was replicated in the laboratory but a number of milling control samples was processed, usually with at least two per day. Trial DH-4 employed partial replication in the laboratory with 0·23 of field plots having two replicates. All field plots from DH-5 were replicated in the laboratory. All milling trials apart from DH-5 were conducted with a fixed number of samples per day. In trial DH-5, the number per day varied between 65 and 80. Full details of the milling layouts are given in Table 4. In terms of randomization of samples for the trials with milling replication, in DH-4 the replicated samples were allocated to days and times within days completely at random. In DH-5 an RCB design was used with a single replicate of all field plots being milled in days 1–5 and then in days 6–10.

STATISTICAL ANALYSIS

In most multi-phase plant breeding experiments, the first phase corresponds to a field trial. It is therefore important to consider issues associated with the analysis of data from field trials.

The literature on methods for the analysis of field trials (which includes the specific setting of variety trials) is quite expansive but the methods can be broadly classified as either randomization- or model-based. In the former, the model for the random and residual effects is determined purely from the block structure, whereas in the latter it is either assumed or selected with the objective of providing a good fit to the data. Model-based approaches for the analysis of variety trials aim to account for the effect of spatial heterogeneity on the prediction of genotype contrasts. Typically, the heterogeneity reflects the fact that, in the absence of design effects (i.e. treatment and block effects), data from plots that are close together (i.e. neighbouring plots) are more similar (positively correlated) than those that are further apart. Numerous authors have proposed analytical methods to remove the effects of such trend. In the present paper the mixed model approach of Gilmour *et al.* (1997) is used. Those authors assume that field trials are arranged as rectangular arrays indexed by rows and columns (extensions to other arrangements are straightforward). Gilmour *et al.* (1997) extended the approach of Cullis & Gleeson (1991) by partitioning spatial variation into two types of smooth trend (local and global) and extraneous variation. Local trend reflects, for example, small-scale soil depth and fertility fluctuations. Global trend reflects nonstationary trend across the field. Extraneous variation is often linked to trial management, in particular, procedures that are aligned with the field rows and columns (e.g. the sowing and harvesting of plots). Certain procedures may result in row and column effects (systematic and/or random). In the Gilmour *et al.* (1997) approach, global trend and extraneous variation are accommodated in the mixed model by including appropriate fixed and/or random effects. Local stationary trend is modelled using a covariance structure for the residuals. A plausible model that has broad application for two-dimensional (row by column) field trials is a separable autoregressive process of order 1 (hereafter denoted AR1 × AR1) as originally proposed by Cullis & Gleeson (1991) and used by Gilmour *et al.* (1997).

The approach presented in the current paper for the analysis of multi-phase plant breeding experiments builds on that of Brien (1983) and Wood *et al.* (1988). In both of those papers the analysis of multi-phase experiments is conducted by determining the experimental structure then including appropriate

Table 4. *DH milling trials: numbers of genotypes, plots (milled either as single samples or replicated) and milling control samples*

	Genotypes		Plots				Layout
Trial	DH	Other	Singles	Rep'd	Total	Milling controls	Days × Times
DH-1	133	4	296	0	296	78	34 × 11 = 374
DH-2	174	8	371	0	371	47	38 × 11 = 418
DH-3	181	2	388	0	388	96	44 × 11 = 484
DH-4	175	8	284	86	370	0	38 × 12 = 456
DH-5	116	4	0	360	360	0	10 × (65–80) = 720

terms in an ANOVA table. The experiments considered by Brien (1983) and Wood *et al*. (1988) are restrictive, however, in the sense that some degree of orthogonality is required. Brien (1983) discusses orthogonal designs that can be analysed using standard ANOVA. Wood *et al*. (1988) consider a class of two-phase designs with non-orthogonal block structure and provide an ANOVA approach to analysis but comment that a full analysis involving recovery of information would require a more sophisticated procedure based on REML estimation of variance parameters. Thus, the linear mixed model provides a natural framework for the analysis of multi-phase experiments. In the present paper, a general linear mixed model that removes any restrictions concerning orthogonality of block structures is proposed. In simple orthogonal cases the approach provides equivalent analyses to those proposed in Brien (1983) and Wood *et al*. (1988).

Hypothetical milling experiment

Before considering this general mixed model, a simple orthogonal two-phase milling experiment that can be analysed using ANOVA is examined. It is assumed that r field replicates of g genotypes are grown in a field trial that is designed as an RCB. Grain samples from each of the rg field plots are split into d smaller samples to be used as replicates in the laboratory process. Thus there is a total of $n = rgd$ samples to be milled. It is assumed that rg samples can be processed each day so that the full trial requires d days. Field plots are randomized to times in the milling process using an RCB design with days as blocks. A single sample from each of the rg field plots is processed each day and the plots are allocated completely at random within a day. The data measured for each sample is the flour yield.

In order to develop the analysis within an ANOVA framework the usual practice of assuming that block effects are random and treatment effects are fixed is followed. Thus, for illustrative purposes it is assumed herein that genotype effects are fixed. As previously noted, the determination of the block structure can be difficult in the context of multi-phase experiments and the concepts in Brien (1983) may be helpful. The basic principle is to include terms in the model that capture the randomization processes used in each phase of the experiment. In the two-phase quality experiment there are two randomizations, namely the randomization of genotypes to field plots then the randomization of field plots to 'positions' in the laboratory process. Thus the effects for block factors associated with each of these randomizations must be included in the analysis. Accordingly, and based on the randomization processes described above, the symbolic model formula for the hypothetical example can be written as

$$y \sim \underline{1} + \underline{\text{genotype}} + \text{mrep} + \text{frep} + \text{frep.plot} + \text{mrep.order} \quad (1)$$

where '1' represents an overall mean, genotype is a factor with g levels, mrep is a factor (for replicates in the milling process) with d levels, frep is a factor (for field replicates) with r levels, plot is a factor (for plots within field replicates) with g levels and order is a factor (indexing the order of processing of samples within days) with rg levels. Thus, the final term in Eqn (1) is the residual term that is also represented generically as units. Note the convention in the model formula of underlining to indicate those terms that correspond to fixed effects. The ANOVA table associated with the model in Eqn (1) is given in Table 5. A key feature of the analysis is the existence of a residual term for each of the two phases. The first phase (field plot) residuals are represented in the model by the term frep.plot and the second phase (laboratory) residuals by mrep.order (or units).

General linear mixed model for multi-phase experiments

The hypothetical milling example is atypical of trials for measuring quality trait data in that the data are rarely balanced, nor are the designs orthogonal. As can be seen from the motivating examples in the

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

Table 5. *ANOVA table for hypothetical milling example*

Strata/Decomposition	D.F.	Model term
mean	1	1
mrep	$d-1$	mrep
mrep.order	$d(rg-1)$	
frep	$r-1$	frep
frep.plot	$r(g-1)$	
genotype	$g-1$	genotype
esidual	$(r-1)(g-1)$	frep.plot
residual	$(d-1)(rg-1)$	units
total	rgd	

section 'Motivating examples', the data-sets are quite complex. Typically, there is partial rather than complete replication in terms of both the field and laboratory. The data from plant breeding trials have an added component of non-orthogonality due to the fact that only a subset of genotypes from the field trial is quality tested. Thus, in general, ANOVA cannot be used but a mixed model analysis must be conducted. However, the same general principles are followed, in particular the inclusion of residual terms for all phases of the experiment.

As with the analysis of field trials, trend in multi-phase trial data can be modelled in order to improve the response to selection. In multi-phase quality trials, the potential exists to model trend (spatial or temporal) associated with the residuals for any of the phases. The type of trend modelling depends on the trait and/or measurement process. Since the present authors' experience has largely been in terms of the analysis of flour yield in wheat and malting quality in barley, modelling is discussed in the context of these data but the concepts generalize to other traits. The modelling approach in the present paper is analogous to that of Gilmour *et al*. (1997) in that trend (either field or laboratory) is partitioned into global trend, extraneous variation and local stationary trend. The latter is accommodated using covariance models. It is important to note that, in the spirit of a randomization-based analysis, terms in the mixed model that are associated with the block structure are maintained irrespective of their level of significance. In contrast, model-based terms and covariance structures are only included if found to be statistically significant.

For simplicity, attention is restricted to two-phase experiments but the extension to more phases is straightforward. An experiment is considered in which the first phase field trial consists of a rectangular array indexed by field rows $(1 \ldots r)$ and columns $(1 \ldots c)$ making a total of $n_p = rc$ plots. In the second phase laboratory trial, it is assumed that a total of n samples is tested. It should be noted that, like the field trial, many laboratory trials can be indexed using a two-dimensional co-ordinate system. For example, in a wheat milling trial the samples may be milled as a set number, s, of samples per day for d days (so that $n = sd$). In a barley malting trial the samples are placed in a machine known as a micro-malter. This has a two dimensional spatial layout of m_r rows and m_c columns so often $n = m_r m_c$. Thus, a rectangular structure for the laboratory process is assumed here. Extensions to non-rectangular or noncontiguous arrays for either the field or laboratory layouts are straightforward. The mixed model for the $n \times 1$ data vector $\boldsymbol{y}$ may be written as

$$\boldsymbol{y} = \boldsymbol{X}\tau + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{Z}_f\boldsymbol{u}_f + \boldsymbol{Z}_b\boldsymbol{u}_b + \boldsymbol{e} \qquad (2)$$

where τ is a $p \times 1$ vector of fixed effects with associated $n \times p$ design matrix $\boldsymbol{X}$ (assumed to have full column rank), $\boldsymbol{u}_g$ is the $g \times 1$ vector of random genotype effects (g is the number of genotypes quality tested) with associated $n \times g$ design matrix $\boldsymbol{Z}_g$, u_b is a $b \times 1$ vector of random block effects with associated $n \times b$ design matrix $\boldsymbol{Z}_b$, $\boldsymbol{u}_f$ is the $n_p \times 1$ vector of random field plot effects with associated $n \times n_p$ design matrix $\boldsymbol{Z}_f$ and $\boldsymbol{e}$ is the vector of residual effects. In the simplest case the fixed effects in Eqn (2) comprise a single effect, namely an overall mean, but may include effects for missing values or covariates to model trend. The vector $\boldsymbol{u}_b$ comprises block effects associated with the experimental design in each phase and other effects as required to model variation. The vector $\boldsymbol{u}_f$ represents the 'error' or residual term for the first (field) phase and $\boldsymbol{e}$ the residual term for the second (laboratory) phase. It should be noted that not all field plots may be quality tested, in which case the matrix $\boldsymbol{Z}_f$ will contain columns whose elements are all zero. Also, note the assumption of random (rather than fixed) genotype effects, which is consistent with the aim of plant breeding experiments: namely the selection of the 'best' genotypes or the identification of QTL (in which case the genotype effects represent residual genetic effects unexplained by the markers).

It is assumed that the joint distribution of $(\boldsymbol{u}_g', \boldsymbol{u}_f', \boldsymbol{u}_b', \boldsymbol{e}')'$ is Gaussian with zero mean and variance matrix

$$\boldsymbol{V} = \sigma^2 \begin{bmatrix} \boldsymbol{G}_g(\boldsymbol{\gamma}_g) & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & G_f(\boldsymbol{\gamma}_f) & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G}_b(\boldsymbol{\gamma}_b) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{R}(\boldsymbol{\phi}) \end{bmatrix}$$

where $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_g', \boldsymbol{\gamma}_f', \boldsymbol{\gamma}_b')'$ is a vector of unknown variance parameters associated with the random effects, $\boldsymbol{\phi}$ is a vector of unknown variance parameters associated with the residuals and σ^2 is the (unknown) scale parameter. The matrix $\boldsymbol{G}_g$ is often a scaled identity matrix, that is, $\boldsymbol{G}_g = \gamma_g \boldsymbol{I}_g$. The associated variance component, $\sigma_g^2 = \sigma^2\gamma_g$, is often termed the genetic variance.

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

Another possibility is $\boldsymbol{G}_g = \gamma_g \boldsymbol{A}$, where $\boldsymbol{A}$ is a known relationship matrix. At present, pedigrees are not generally used in routine analyses of variety trials so will not be considered in the present paper. The matrix $\boldsymbol{G}_b$ is usually a direct sum of scaled identity matrices, each component corresponding to different terms within $\boldsymbol{u}_b$. Forms for $\boldsymbol{G}_f$ and $\boldsymbol{R}$ are discussed below.

As an illustration, consider the ANOVA for the hypothetical milling experiment discussed earlier, but now regard the genotype effects as random rather than fixed. The vector $\boldsymbol{u}_g$ in Eqn (2) corresponds to the term genotype in Eqn (1); $\boldsymbol{u}_f$ corresponds to frep.plot; $\boldsymbol{u}_b$ contains sub-vectors corresponding to mrep and frep; $\boldsymbol{e}$ corresponds to mrep.order (or units) and τ contains a single parameter only, namely an overall mean. In the ANOVA model, each set of random effects is assumed to be independent and each has an associated variance component so that $\boldsymbol{G}_g = \gamma_g \boldsymbol{I}_g$, $\boldsymbol{G}_f = \gamma_f \boldsymbol{I}_{rg}$, $\boldsymbol{G}_b = diag(\gamma_{m_r} \boldsymbol{I}_d,\ \gamma_{f_r} \boldsymbol{I}_r)$ and $\boldsymbol{R} = \boldsymbol{I}_n$. Thus, the general mixed model of Eqn (2) encompasses ANOVA models for multi-phase data (as discussed in Brien 1983; Wood *et al.* 1988, for example) but has much broader application since non-orthogonal designs and unbalanced data are easily handled and more general covariance structures can be considered. The latter is particularly important for the modelling of local stationary trend associated with either the field or laboratory phase.

Covariance models for stationary trend

In terms of covariance models for the field residuals, the approach is as for the spatial analysis of a field trial. The vector of field residuals is assumed ordered as field rows within columns so that

$$\boldsymbol{G}_f = \boldsymbol{G}_c(\boldsymbol{\gamma}_c) \otimes \boldsymbol{G}_r(\boldsymbol{\gamma}_r)$$

where $\boldsymbol{G}_c$ is the $c \times c$ correlation matrix for columns and $\boldsymbol{G}_r$ is the $r \times r$ matrix for rows and $\boldsymbol{\gamma}_c$ and $\boldsymbol{\gamma}_r$ are vectors of unknown parameters. As previously discussed, autoregressive processes of order one provide plausible models for $\boldsymbol{G}_c$ and $\boldsymbol{G}_r$. The associated variance parameters are simply ρ_c and ρ_r (so that $\boldsymbol{\gamma}_f = (\rho_c, \rho_r)'$) and these are known as the autocorrelation parameters for columns and rows respectively.

In terms of the laboratory phase, first consider flour yield data. Recall that the milling of samples from a field trial involves a sequential process that usually requires more than a single day. There is potential for temporal correlation linked to the order of processing samples within a day. If a rectangular trial layout is assumed, with d days and s samples per day (making a total of $n = ds$ samples) and the data are ordered sequentially within days, then the temporal correlation leads to a correlation matrix for the residuals of the form

$$\boldsymbol{R} = \boldsymbol{I}_d \otimes \boldsymbol{R}_o(\boldsymbol{\rho}_o)$$

where $\boldsymbol{R}_o$ is the $s \times s$ correlation matrix for sample order within days. As in the spatial modelling of field trials, a range of covariance models is possible. The present authors have found that an autoregressive process of order 1 provides a plausible model for $\boldsymbol{R}_o$. The full correlation model for $\boldsymbol{e}$ is therefore denoted by ID $\times$ AR1. The associated autocorrelation parameter is denoted ρ_o so that $\phi = \rho_o$.

In terms of the measurement of malting quality, recall that grain samples are tested in a micro-malter machine. Due to the arrangement of samples in the micro-malter, there is potential for spatial variation. Let m_r and m_c denote the numbers of rows and columns in the micro-malter. For simplicity it is initially assumed that all samples can be processed in a single 'run' of the micro-malter so that $n = m_r m_c$. If the data are ordered as micro-malter rows within columns then

$$\boldsymbol{R} = \boldsymbol{R}_{m_c}(\boldsymbol{\rho}_{m_c}) \otimes \boldsymbol{R}_{m_r}(\boldsymbol{\rho}_{m_r})$$

where $\boldsymbol{R}_{m_r}$ and $\boldsymbol{R}_{m_c}$ are the correlation matrices for the micro-malter row and column dimensions. Once again, the present authors have found the AR1 $\times$ AR1 model to be reasonable so that $\boldsymbol{\phi} = (\rho_{m_c},\ \rho_{m_r})'$. If the data comprises several runs of the micro-malter then independence of the errors between runs is assumed and often the autocorrelation parameters are constrained to be the same for all runs.

Estimation and inference

Estimation of a linear mixed model involves two processes. First the variance parameters ($\boldsymbol{\gamma}$, $\boldsymbol{\phi}$ and σ^2) are estimated, with residual maximum likelihood (REML, Patterson & Thompson 1971) being the preferred method. Given estimates of the variance parameters, the fixed effects in the model are estimated using empirical best linear unbiased estimation (E-BLUE) and the random effects predicted using empirical best linear unbiased prediction (E-BLUP). In particular E-BLUPs for the genotype effects are obtained, denoted by $\tilde{\boldsymbol{u}}_g$, and these are the basis of selection decisions. In addition, an approximate prediction error variance matrix for the genotype effects, denoted $\boldsymbol{V}_{gg}$, is obtained. Cullis *et al.* (in press) show how this matrix may be used to calculate heritability and Expected Genetic Gain (EGG) for complex datasets such as those generated in multi-phase trials. Cullis *et al.* (in press) obtain a generalized measure of heritability as

$$h_g^2 = 1 - \frac{a}{2\sigma^2 \gamma_g} \tag{3}$$

where a is the average pairwise prediction error variance of genotype effects, that is

$$a = \frac{2}{g-1}\left(\mathrm{tr}(\boldsymbol{V}_{gg}) - \frac{1}{g}\boldsymbol{1}_g' \boldsymbol{V}_{gg} \boldsymbol{1}_g\right)$$

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

Note that $\mathbf{1}_g' \boldsymbol{V}_{gg} \mathbf{1}_g = 0$ unless selection (and hence the heritability calculation) is limited to a subset of the genotypes (for example, if the genotypes comprise test and standard lines with the latter being excluded from the selection process). The expected genetic gain from selection of the top m genotypes is then calculated as

$$\mathrm{EGG} = i\sqrt{\sigma^2 \gamma_g h_g^2} \tag{4}$$

where i is the 'selection intensity' corresponding to m (that is, the mean of the top m order statistics from a standard normal distribution of size g). Note that in the present paper a key element is the impact of statistical modelling on genetic gain. Modelling may increase h_g^2 (and thence genetic gain as shown in Eqn (4)) by causing a reduction in error variance (or effective error variance for those analyses, including the analysis of multi-phase data, that do not include an explicit error variance).

The significance of fixed effects in the mixed model may be assessed using Wald tests. These are asymptotic tests that may be anti-conservative for small samples sizes. In these cases, the adjustments of Kenward & Roger (1997) may be used. Nested variance models may be compared using residual maximum likelihood ratio tests (REMLRT). If the test involves a null hypothesis with a parameter on the boundary of the parameter space (e.g. a test of a zero value for a variance component when the component has been constrained to be non-negative), then an adjustment is required for the significance level (see Stram & Lee 1994, for example).

All analyses in the present paper were conducted using either ASReml (Gilmour *et al.* 2002) an efficient program for fitting complex mixed models or the samm (Butler *et al.* 2003) suite of functions (written for use within S-language environments; Becker *et al.* 1988) that uses the core routines of ASReml.

Analysis of example data-sets

First, a detailed account of the analysis of the flour yield data from trial DH-5 is presented (see Tables 3 and 4 for trial details). Note that these data do not comprise a rectangular array since the number of samples milled each day was not constant (varying from 65 to 80). For simplicity, the data-set has been expanded from 720 to 800 observations indexed as 10 days by 80 samples per day so that a rectangular layout is achieved. This is not necessary, but allows the computational advantages of separability for the residual variance structure to be exploited. The 80 additional observations have a missing value indicator for the dependent variable.

The analysis commenced by considering the base-line mixed model that contains random effects for all block terms. The design for this trial involved an RCB in both the field and laboratory phases so that the model can be written in symbolic notation as

$$y \sim \underline{1} + \underline{\mathrm{gfac}} + \mathrm{genotype} + \mathrm{mrep} + \mathrm{frep} + \mathrm{column.row} + \mathrm{day.order} \tag{5}.$$

where gfac is a factor with 5 levels (levels 1 to 4 correspond to the parental genotypes and commercial varieties and level 5 is assigned to all DH genotypes), genotype is a factor with 120 levels, frep is a factor with 3 levels (indexing field replicate blocks), column and row are factors with 12 and 30 levels indexing field columns and rows respectively, mrep is a factor with 2 levels (indexing milling replicate blocks), day and order are factors with 10 and 80 levels indexing milling days and order within days, respectively. All effects in the model, apart from the overall mean and gfac, are fitted as random effects. The fixed term gfac is included in order to ensure that the variance component associated with the genotype effects relates to the DH genotypes alone and not all the genotypes in the trial. Note that the terms column.row and day.order represent the field and laboratory residuals respectively (i.e. $\boldsymbol{u}_f$ and $\boldsymbol{e}$ in Eqn (2)). Initially, these were assumed to represent independent random variables each with constant variance, denoted by $\sigma_f^2 = \sigma^2 \gamma_f$ and σ^2 respectively. The term genotype corresponds to $\boldsymbol{u}_g$ in Eqn (2) and its variance is denoted by $\sigma_g^2 = \sigma^2 \gamma_g$; the terms frep and mrep represent subvectors in $\boldsymbol{u}_b$ with variance components $\sigma_{f_r}^2 = \sigma^2 \gamma_{f_r}$ and $\sigma_{m_r}^2 = \sigma^2 \gamma_{m_r}$. Note that the model in Eqn (5) is almost identical to that in Eqn (1) for the hypothetical milling experiment, since the same types of randomizations were employed. One difference here is that the field plot residual term has been indexed using column and row numbers rather than replicate and plot within replicate numbers since this is required for spatial modelling purposes.

The estimated variance components from the fit of the base-line model are given in Table 6 under the column headed 'M0'. Note that the genetic variance represented a large proportion (0·84) of the total variation in the data. This was expected due to the nature of the DH population. A QQ-plot (Wilk & Gnanadeskin 1968) of the estimated laboratory residuals from this model revealed some potential outliers (Fig. 1). The outliers related to discrepancies in the two laboratory replicate values for three field plots. The very low flour yield values for plots (column 1, row 22 and column 1, row 26) and the high value for plot (column 3, row 26) were found to be erroneous so were omitted from subsequent analyses.

The base-line mixed model was re-fitted and the resultant QQ-plot for the laboratory residuals was satisfactory. The estimated variance components from this model are given under the heading 'M1' in Table 6. Note that the identification and exclusion of the three outliers resulted in a substantial reduction

Table 6. *Estimates of variance parameters and key fixed effects in analysis of DH-5 milling trial*

Model term	Parameter	M0	M1	M2	M3	M4	M5
genotype	σ_g^2	2·293	2·330	2·316	2·285	2·289	2·276
frep	$\sigma_{f_r}^2$	0·006	0·003	0·003	0·004	0·004	0·002
column.row	σ_f^2	0·141	0·122	0·128	0·124	0·101	0·104
	ρ_c						0·408
	ρ_r						0·263
mrep	$\sigma_{m_r}^2$	0·014	0·019	0·019	0·018	0·018	0·018
units	σ^2	0·280	0·153	0·142	0·148	0·147	0·149
	ρ_o				0·569	0·560	0·563
lin(order) × 10^2	τ_o			−0·1	−0·1	−0·1	−0·1
lin(row) × 10^2	τ_r					−1·7	−1·7
Residual log-likelihood		−215·26	−77·25	−73·75	−29·30	−19·29	−14·32
Effective error variance		0·622	0·442	0·443	0·382	0·382	0·380

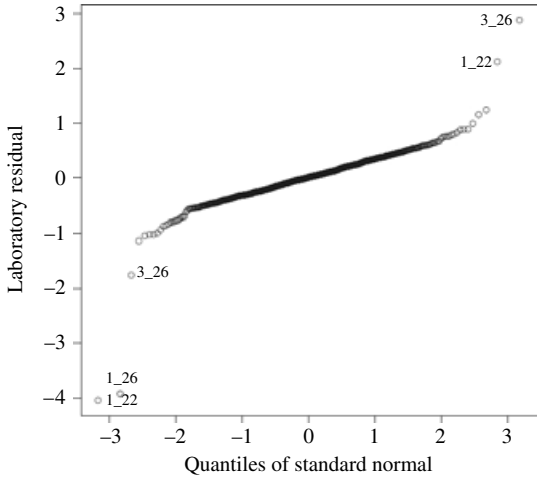

Fig. 1. Analysis of DH-5 trial: QQ-plot for estimated laboratory residuals from model M0. Potential outliers are labelled with their field column and row numbers.

in residual variance and also effective error variance (0·622 for model 'M0' compared with 0·442 for model 'M1'). In terms of non-genetic variance, both the field and laboratory made substantial contributions (0·42 and 0·58 of the total non-genetic variance, respectively). The plot of laboratory residuals from model 'M1' against milling order for each day (Fig. 2) suggested the existence of global trend, namely a linear decline in flour yield over the course of each day. Thus, a linear regression on milling order was added to the model and found to be significant ($P<0·001$). Having accommodated this global trend across milling order within a day, an AR1 correlation structure was added to accommodate local trend. The estimated autocorrelation parameter was 0·569 (see model 'M3' in Table 6) and is significant ($\chi_1^2=88·90$, $P<0·001$).

Having established a plausible model for laboratory trend, the modelling of field trend was then considered. This is possible due to the large contribution of field trend to total non-genetic variance. The graph of field plot residuals from model 'M3' against field row number for each column (Fig. 3) suggested the existence of global trend in the form of a linear decline in milling yield over row number. Thus, a linear regression on row number was added to the model and found to be significant ($P<0·001$). Finally, an AR1 × AR1 correlation structure for the field residuals was added. The estimated autocorrelation parameters were 0·41 and 0·26 for the column and row dimensions, respectively (see model 'M5' in Table 6) and are significant ($\chi_2^2=9·96$, $P<0·001$). Note that the modelling process facilitated a reduction in effective error variance from 0·442 for model 'M1' to 0·380 for model 'M5'.

Flour yield data for the remaining nine examples described earlier were analysed using the same approach as for trial DH-5. The resultant estimates of variance parameters and key fixed effects from the final models for each data-set are summarized in Table 7. A key feature of the table is the consistency of trend observed in the laboratory phase. The trend is usually manifested both as global trend (with a linear decline in flour yield from the beginning to the end of each day) and as local stationary trend with strong correlation (autocorrelation parameters ranging from 0·56 to 0·82). The consistency is particularly noteworthy given that three different laboratories and two different milling methods were used in the present set of 10 trials.

Another interesting feature is that in terms of error variation, the contribution from the first (field) phase was generally lower than from the second (laboratory) phase. This is consistent with the present authors' experience in analysing flour yield data. The spatial modelling of field trend was only possible in

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

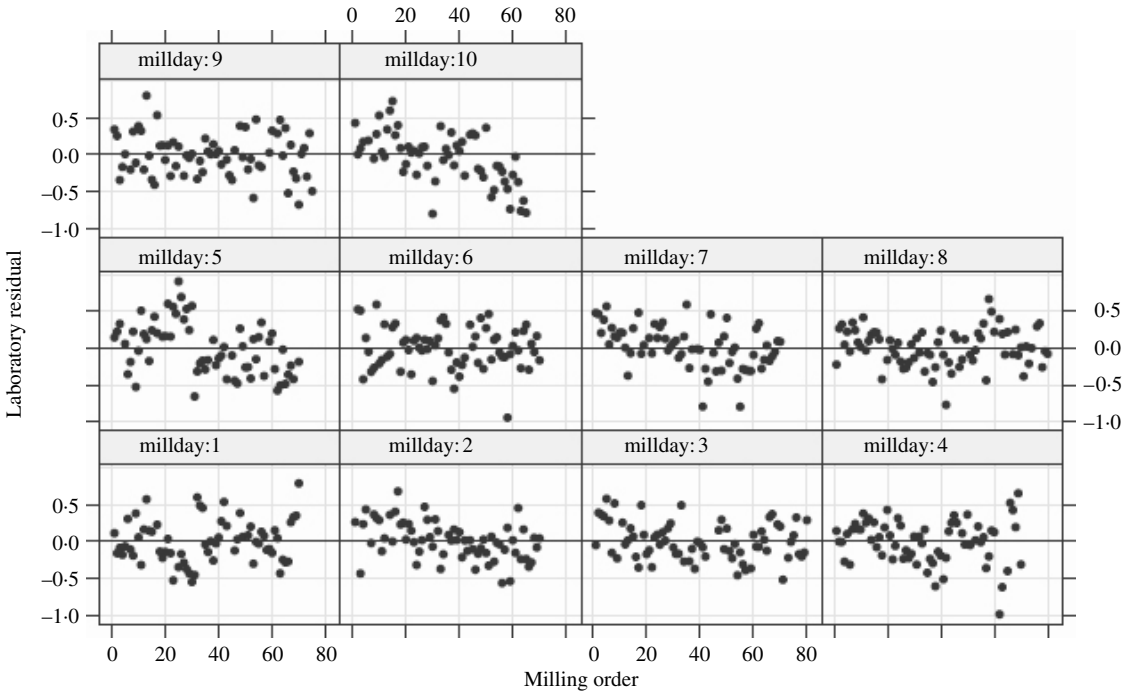

Fig. 2. Analysis of DH-5 trial: estimated laboratory residuals from model M1 graphed against milling order for each milling day.

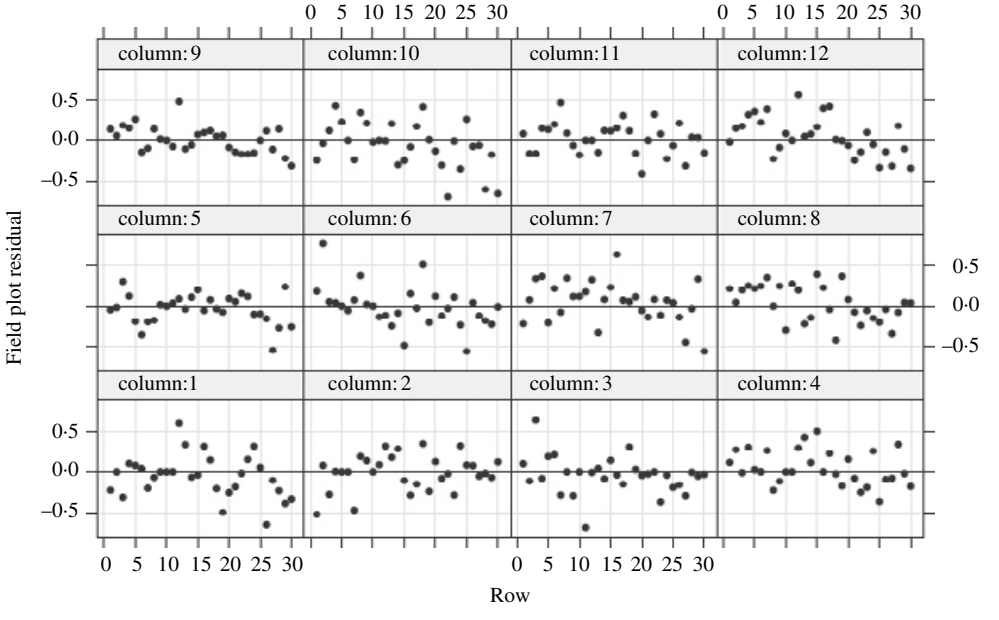

Fig. 3. Analysis of DH-5 trial: estimated field plot residuals from model M3 graphed against row number for each column.

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

Table 7. *Analysis of ten flour yield data-sets: estimated variance parameters (components unless otherwise indicated) and fixed effects for final models. Model terms partitioned into block terms and model based terms. Estimated variance components are given for all block terms unless the term was not applicable for a particular design (so left blank). Estimated parameters for model-based terms are only given if parameter was significant at the 5% level (otherwise left blank)*

	Block terms						Model-based terms				
Trial	geno	frep	col.row	mrep	day	units	lin(row) × 10²	ρ_c	ρ_r	lin(ord) × 10²	ρ_o
NSW-1	1·39		0	0·193	0	0·726				−3·3	0·78
NSW-2	1·36		0·259	0·039	0	0·271				−3·8	0·82
NSW-3	0·98		0·008	0·124	0·104	0·437				−2·7	
NSW-4	1·88		0·120	0	0·309	0·318				−2·0	0·61
NSW-5	0·91		0·302	0·671	0·960	0·600				−0·9	0·66
DH-1	0·87	0	0		0·329*	0·213				−5·7	
DH-2	2·81	0·133	0·340			0·784		0·70	0·84	−6·3	0·71
DH-3	1·44	0	0·617		0·225*	0·201		0·91	0·99	−3·7	
DH-4	1·92	0·077	0·406		0·482*	0·271	−2·1			−4·4	0·71
DH-5	2·28	0·002	0·104	0·018		0·149	−1·7	0·41	0·26	−0·1	0·56

* Term was not part of block structure but was a significant model-based term.

three of the trials and all of these were DH trials. In terms of genetic variance, it should be noted that the estimates for the NSW trials are representative of early generation selection trials and the estimates for the DH trials are substantially larger since they reflect populations constructed to be variable for this trait.

A NEW CLASS OF EXPERIMENTAL DESIGNS

As is evidenced from the analysis of the motivating examples, multi-phase quality trait data may exhibit 'nuisance' trend arising from the field and/or laboratory phases. The potential exists to increase response to selection by appropriate statistical modelling of this trend. This can only be achieved with the use of an experimental design that employs replication and randomization in all phases. In the specific application of quality trait testing, it is important to consider the high cost associated with the acquisition of the data and the consequent need to limit the total number of samples tested for any given field trial. Fully replicated designs are prohibitively expensive and are unnecessary from a statistical perspective. In the present paper, a scheme based on the staggered nested designs of Bainbridge (1965) and *p*–rep designs of Cullis *et al.* (in press) is proposed. This is illustrated by means of a simple example in which 40 genotypes are grown in a field trial with 2 replicates, then all genotypes are to be milled in the laboratory. A fully replicated design (assuming two laboratory replicates) would require 160 samples to be milled (that is, 40 genotypes × two field replicates × two laboratory replicates). This scheme is shown diagramatically in Table 8 (*a*). The design is balanced for genotypes, in the sense that there is an equal number (four) of observations for each genotype. This scheme uses a large amount of resources and is 'bottom' heavy in terms of degrees of freedom (D.F.) for estimating error, with 40 D.F. for plot error and 80 D.F. for laboratory error. Bainbridge (1965) devised staggered nested designs in order to overcome this problem. In the context of the example, one field replicate of each genotype is replicated in the laboratory whilst the other is only tested once, making a total of 120 samples. This design is also balanced for genotypes, with three observations for each genotype. This scheme is shown diagrammatically in Table 8 (*b*). The D.F. for estimating plot and laboratory error are both 40 and the scheme requires considerably less resources than the fully replicated design. The present authors believe that further economies are possible using the partial replication idea of Cullis *et al.* (in press). This involves the use of field replicates for a proportion p of the genotypes and laboratory replicates for a proportion q of the field plots. In order to equalize, as far as possible, the number of observations for each genotype it is ensured that the field plots that are replicated in the laboratory correspond to genotypes for which only a single field replicate is tested. One such scheme for the example under consideration is illustrated in Table 8 (*c*). In this scheme $p = 0{\cdot}25$ and $q = 0{\cdot}20$. The total number of samples is only 60 and the number of observations for each genotype is either 1 or 2. The D.F. for estimating plot and laboratory error are both 10. These designs, that shall be called p/q–rep designs, clearly have much to offer in the costly area of quality testing, in which the total number of samples must be kept to a minimum. The

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

Table 8. *Diagrammatic representation of example laboratory designs for testing 40 genotypes from a 2 replicate field trial and with 2 laboratory replicates. (a) Fully replicated design. (b) Staggered nested design. (c) p/q design with p = 0·25 and q = 0·20*

(*a*)

	Genotypes	Field plots	Milling samples
	40	40	40
			40
		40	40
			40
Total	40	80	160

(*b*)

	Genotypes	Field plots	Milling samples
	40	40	40
			40
		40	40
Total	40	80	120

(*c*)

	Genotypes	Field plots	Milling samples
	10	10	10
		10	10
	10	10	10
			10
	20	20	20
Total	40	50	60

extension to experiments with more than two phases is obvious.

Assessing performance of new designs

It was hypothesized that the ability to model trend facilitated with the use of a *p*/*q*–rep design would, for a given total number of samples, lead to higher genetic gains than with the use of a 'design' without any replication. This could be investigated using an algebraic approach similar to that of Kempton (1984). Thus, Eqn (4) could be used to establish the reduction in effective error variance that would be required to equalize the genetic gain associated with the two designs. A key issue, however, is that Eqn (4) is based on the premise that variance parameters are known, whereas in practice they must be estimated from the mixed model analysis. With the proposed designs the level of replication may be quite low; therefore knowledge of how this may adversely affect variance parameter estimates is important. Thus, a simulation study was conducted in order to calculate realized genetic gain (based on E-BLUPs from a mixed model analysis) and use this, rather than expected genetic gain, as a basis for comparing designs.

Table 9. *Designs compared in simulation study for testing genotypes from a field trial with 576 genotypes and 720 plots in a laboratory trial with 448 samples. Total number of genotypes tested for each design is: 448 for D00 and Dc0; 407 for Dcq and 370 for Dpq*

Design	Genotypes	Field plots	Milling samples
D00	448	448	448
Dc0	112	112	112
		112	
Dcq	336	336	336
	102	102	102
		102	
Dpq	41	41	41
			41
	264	264	264
	37	37	37
		37	37
	41	41	41
			41
	292	292	292

Under consideration were four designs for a milling experiment that is typical of an early generation trial in the NSWDPI wheat breeding programme. A field trial comprising 720 plots (arranged as 12 columns × 60 rows) and a total of 576 genotypes is assumed. The field trial uses a *p*–rep design (Cullis *et al*., in press) with $p = 0{\cdot}25$ so that 144 genotypes are replicated and the remaining 432 are planted as single plots. It is then assumed that a fixed number (448) of samples can be milled in the laboratory and that it is possible to mill 28 samples per day so that the complete milling will require 16 days.

The designs compared comprised a 'null' design (that shall be denoted D00), in which there is no replication in either phase so that a total of 448 genotypes may be tested (with a single field plot of each). Also considered was the scenario in which the same 448 genotypes were tested, but for those genotypes with two field replicates the data were averaged to mimic a composite field sample. This design is denoted by Dc0. The D00 and Dc0 designs were, until recently, used quite commonly for milling trials and remain the standard approach for many other quality traits. The third design considered, labelled Dcq, was similar to Dc0 in that composite field samples were used but replication in the laboratory was included with $q = 0{\cdot}10$ of field samples replicated. This allowed a total of 407 genotypes to be tested. The final design was a *p*/*q*–rep design with $p = q = 0{\cdot}10$. This allowed 370 genotypes to be tested. Details of the replication involved in all four designs are given in Table 9.

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

In terms of the randomization of samples in the milling process, the samples for D00 were processed in field order (rows within columns) and the samples for Dc0 were processed in the same order. The samples for Dcq and Dpq were randomized, with the samples involving laboratory replication being allocated to positions in the milling process using a resolvable incomplete block design with days as blocks and the remaining (unreplicated) samples were allocated at random to the 'vacant' positions (also see description of milling designs for the NSW trials as given in the section 'Motivating examples'). The designs employing replication were generated using the DiGGer software (Coombes 2002).

Data for each design were generated according to 21 different models that were chosen as being typical of flour yield data. Due to the consistent occurrence of laboratory trend (see 'Analysis of example datasets'), each data model included a (fixed) linear regression on sample order (slope = –0·03), random effects for milling days (variance component = 0·32) and an autoregressive process over sample order (autocorrelation parameter = 0·60). The residual variance (scale parameter) was taken to be $\sigma^2 = 0{\cdot}38$. All models could then be represented symbolically as

$$y \sim \underline{1} + \underline{\text{lin(order)}} + \text{genotype} + \text{column.row} + \text{day} + \text{day.order} \quad (6)$$

In terms of genetic effects, three different variance component ratios were used, namely $\gamma_g = 0{\cdot}5$, 1·0 and 2·0. In terms of field plot effects, seven different models were used, namely no plot effects (that is, a variance component ratio for the term column.row of $\gamma_f = 0$) and the factorial combination of three non-zero values of γ_f (0·5, 1·0 and 2·0) by two correlation models (independence and a separable autoregressive process of order 1 with correlation parameters of $\rho_c = 0{\cdot}4$ and $\rho_r = 0{\cdot}6$).

In accordance with the two-phase nature of the experiment, the data were generated in two stages. Non-genetic effects associated with the laboratory (linear regression on order, random day effects and residuals) were generated first since, for any given simulation, these remained constant across all 21 data models. Then genetic and field plot effects were generated. This was done in reference to the complete field trial even though only a subset was tested in the laboratory. Thus, 576 genetic effects and 720 plot effects were generated. The four designs involve different numbers of genotypes and field plots; therefore the next step was to extract the appropriate subsets of effects for each design. In order to improve the accuracy of comparisons of designs, this was carried out in a nested manner. Thus, in terms of the genotype effects, the same 448 genotypes were used for D00 and Dc0, the 407 genotypes used in Dcq were a subset of those used in D00 and Dc0 and the 370 genotypes used in Dpq were a subset of those used in Dcq. An analogous approach was used for the field plot effects. In any of the designs, the genotypes comprised a mixture of those that had been replicated in the field and those that were grown as single plots. The ratio was kept constant in all designs and was set at 1:3 to match the original field trial in which 0·25 of genotypes were replicated.

Note that the data model does not include terms for resolvable blocks in either the field or laboratory. In practice these terms would be included in order to respect the randomization, but in terms of the simulation they add complexity since a range of values may need to be used. Also, resolvable blocks are not part of the design model but are accommodated using a restriction of the search algorithm. For these reasons, and without loss of generality, the effects have been excluded from the data model.

Simulated data were analysed using models that matched the data generation model, subject to restrictions imposed by the design. Thus, for the designs D00 and Dc0, there was no analysis since there was no replication in the data. Data for Dcq was analysed using the model in Eqn (6), except that the term column.row was omitted (since field plot replicates were not maintained but combined to form composite samples). Finally, data for Dpq were analysed using the full model as in Eqn (6).

A total of 200 simulations was conducted for each combination of 21 data models and four designs. In each simulation, the true and realized genetic gain (denoted TGG and RGG) were calculated for the selection of 70 genotypes. TGG was calculated for all designs by ranking the genotypes on the basis of the true genetic effects, selecting the top 70 and calculating their mean. RGG for designs D00 and Dc0 was calculated by ranking the genotypes on the basis of the raw data, selecting the top 70 then calculating the mean of the associated true genetic effects. RGG for Dcq and Dpq was calculated in a similar way, but genotypes were ranked on the basis of the E-BLUPs from the mixed model analysis. TGG is only influenced by the design (due to differences in number of genotypes tested and thence proportion selected) and the genetic variance. The average TGG values for the four designs and three levels of genetic variance are given in Table 10. When scaled by the genetic standard deviation $\left(\sqrt{\sigma^2\gamma_g}\right)$ the average (standardized) TGG for each design is equivalent to the intensity of selection (see final column in Table 10). The selection intensity for designs D00 and Dc0 are identical and the intensity is lowest for Dpq, which encompasses the most replication (and thence least number of genotypes tested). Equation (4) clearly shows that, for a given trial size and genetic variance, a decrease in intensity must cause a decrease in genetic gain unless the effective error variance can be reduced (and thence heritability increased). Thus, in terms of RGG,

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

Table 10. *Simulation study: mean over 200 simulations of True Genetic Gain (TGG) for selection of 70 genotypes for four designs and three values of genetic variance. Final column is average standardized TGG (that is, intensity of selection) across all values of genetic variance*

Design	Genotypes tested	Proportion selected	TGG			Intensity of selection
			$\gamma_g = 0{\cdot}5$	$\gamma_g = 1{\cdot}0$	$\gamma_g = 2{\cdot}0$	
D00	448	0·16	0·666	0·941	1·333	1·528
Dc0	448	0·16	0·666	0·941	1·333	1·528
Dcq	407	0·17	0·644	0·910	1·289	1·477
Dpq	370	0.19	0·621	0·879	1·244	1·426

the issue is whether the statistical modelling associated with the designs that employ replication, in particular the *p*/*q*–rep design (Dpq), has a large enough impact on effective error variance to offset the reduction in intensity as quantified in Table 10.

The mean RGG across 200 simulations for the four designs and 21 data models described in 'Assessing the performance of new designs' is given in Table 11. The most obvious feature of this table is that RGG for the designs that use laboratory replicates (Dcq and Dpq) is always higher than for the designs that do not (D00 and Dc0). Thus, the modelling of laboratory trend has facilitated a reduction in effective error variance that has outweighed the reduction in selection intensity induced by testing fewer genotypes. The impact of testing individual field replicates is less general. The scheme that employs composite field samples (Dcq) resulted in higher RGG than the *p*/*q*–rep design for 16 of the 21 data models. This included all data models in which there was no field spatial correlation. When there was spatial correlation in the field and the associated variance was relatively large ($\gamma_f = 2{\cdot}0$) the *p*/*q*–rep design outperformed the Dcq design, with larger benefit for smaller genetic variance. Note that in terms of the ability of each of the four designs to achieve the potential genetic gains, RGG can be considered as a percentage of TGG. Once again, across all 21 data models Dcq and Dpq were substantially better than D00 and Dc0. Additionally, Dpq was always the same as or better than Dcq. The average values of RGG as a proportion of TGG for the four designs are 0·51, 0·52, 0·64 and 0·65 for D00, Dc0, Dcq and Dpq, respectively. It is also important to note that the variability of RGG values was substantially lower for the designs with laboratory replication and, on average, was lower for the *p*/*q*–rep design compared with the Dcq design (Table 12).

Table 11. *Simulation study: mean Realized Genetic Gain (RGG) for selection of 70 genotypes for four designs and 21 data models. RGG for design D00 is given in units of measurement; RGG for other designs are expressed as a proportion of D00. n = 200 simulations*

Data model			Design			
Field spatial	γ_g	γ_f	D00	Dc0	Dcq	Dpq
No	0·5	0	0·304	1·014	1·450	1·436
No	0·5	0·5	0·273	1·031	1·327	1·277
No	0·5	1	0·250	1·051	1·284	1·222
No	0·5	2	0·220	1·041	1·199	1·123
No	1	0	0·553	1·000	1·282	1·261
No	1	0·5	0·515	0·991	1·217	1·189
No	1	1	0·478	1·016	1·197	1·153
No	1	2	0·418	1·021	1·153	1·117
No	2	0	0·952	0·994	1·157	1·134
No	2	0·5	0·898	1·011	1·139	1·107
No	2	1	0·853	1·021	1·123	1·079
No	2	2	0·763	1·034	1·117	1·073
Yes	0·5	0·5	0·279	1·009	1·296	1·297
Yes	0·5	1	0·255	1·027	1·243	1·272
Yes	0·5	2	0·219	1·038	1·206	1·282
Yes	1	0·5	0·508	1·022	1·239	1·222
Yes	1	1	0·469	1·025	1·203	1·200
Yes	1	2	0·432	1·031	1·135	1·181
Yes	2	0·5	0·887	1·008	1·148	1·117
Yes	2	1	0·854	1·012	1·119	1·098
Yes	2	2	0·780	1·025	1·102	1·114

DISCUSSION

In the present paper, a mixed model approach for the analysis of multi-phase quality trait data and a new class of designs that employs partial replication in all phases has been described. In terms of analysis, the proposed model generalizes the work of Brien (1983) and Wood *et al.* (1988). Those authors use analysis of variance tables that include all sources of variation as necessary to represent the block structure in each phase. This principle is followed in the present paper, but a modelling aspect is added in order to accommodate additional sources of variation and correlation. The approach easily handles non-orthogonal designs and unbalanced data, which tend to be the

https://doi.org/10.1017/S0021859606006319 Published online by Cambridge University Press

Table 12. *Simulation study: standard error of mean Realized Genetic Gain (RGG) for selection of 70 genotypes for four designs and 21 data models. Standard error for design D00 is given in units of measurement × 100; standard errors for other designs are expressed as a proportion of D00. n = 200 simulations*

Data model			Design			
Field spatial	γ_g	γ_f	D00	Dc0	Dcq	Dpq
No	0·5	0	0·396	0·981	0·814	0·798
No	0·5	0·5	0·383	0·980	0·925	0·876
No	0·5	1	0·398	0·944	0·864	0·932
No	0·5	2	0·380	0·924	0·950	0·957
No	1	0	0·591	0·940	0·728	0·728
No	1	0·5	0·488	1·072	0·906	0·871
No	1	1	0·500	0·954	0·968	0·974
No	1	2	0·523	1·059	0·954	1·011
No	2	0	0·705	1·002	0·780	0·807
No	2	0·5	0·678	0·915	0·903	0·830
No	2	1	0·721	0·945	0·898	0·786
No	2	2	0·670	1·036	1·025	0·993
Yes	0·5	0·5	0·394	0·895	0·840	0·885
Yes	0·5	1	0·362	0·951	0·944	0·861
Yes	0·5	2	0·378	0·992	0·917	0·916
Yes	1	0·5	0·513	0·962	0·900	0·872
Yes	1	1	0·509	1·021	0·892	0·928
Yes	1	2	0·510	1·013	1·036	0·880
Yes	2	0·5	0·732	0·967	0·808	0·870
Yes	2	1	0·753	0·932	0·857	0·878
Yes	2	2	0·676	1·022	1·020	0·914

rule rather than the exception in the case of quality trait data. The flexibility of the mixed model was illustrated with the analysis of 10 wheat flour yield data-sets. Most of these data were highly unbalanced. The modelling of trend was an important aspect for all data-sets, resulting in substantial reductions in effective error variance. This phenomenon underpins the proposed approach to experimental design.

The basic principle of the new designs for a two-phase quality experiment is to test field replicates of a proportion p of genotypes (and use single plots of the remainder) and test laboratory replicates of a proportion q of the field plots (and use single samples of the remainder). This approach has the benefit of facilitating the use of the mixed model analysis, thence providing a reduction in effective error variance that is likely to lead to an increase in genetic gain. At the same time, the total number of samples is limited, thence controlling the overall cost of testing. The latter is a key issue for Australian plant breeding programmes.

A simulation study was conducted to compare the new designs (so-called p/q–rep designs) with schemes that are commonly used in two-phase quality testing. Thus, the scheme with no replication in either phase was considered, as was a scheme that had no replication in the laboratory but used composite field samples. For completeness, a design that used composite field samples and then laboratory replication for a proportion of these was also considered. Note that another design in common usage, namely a design in which multiple control samples are tested in the laboratory process, was not considered, since this scenario is analogous to the grid-plot field design in which multiple control genotypes are grown. Cullis *et al.* (in press) show that the genetic gain associated with their partially replicated (p–rep) field designs is superior to that for grid-plot designs, so the same result is inferred in the present paper in terms of the laboratory design in a multi-phase experiment.

The simulation study was based on data models typical of flour yield data from wheat breeding trials. The study showed that, for a fixed total number of samples tested, the p/q–rep design with $p = q = 0{\cdot}10$ was superior to the no replication scheme in terms of realized genetic gain. The average gain (across all data models) for the p/q–rep design was 19% higher than that for the no replication scheme. Thus, the modelling of trend facilitated with the use of the p/q–rep design achieved sufficiently large reductions in error variance to offset the lower selection intensity compared with the no replication design. The overall impact of modelling can be partitioned into the components associated with the field and the laboratory. The study showed that the designs that used laboratory replication always had higher realized genetic gain than those that did not. The conclusion from this is that, in order to maximize genetic gain, laboratory replication is essential for quality traits that exhibit trend in the laboratory phase. Wheat flour yield is one such trait and the present authors' experience has been that a number of other traits, including barley malting quality traits (see Cullis *et al.* 2003) also exhibit substantial trend in the laboratory process.

In terms of the testing of field replicates, the present study showed that when there was no spatial correlation in the field, the realized genetic gain for the p/q–rep design (which processes individual field replicates) was lower than that for the design in which composites of replicates are tested. However, in the presence of spatial correlation with a relatively large variance and a genetic variance the same size or smaller than residual variance, the genetic gain for the p/q–rep design was substantially higher than for the design using composite samples. In terms of the trait studied here, namely wheat flour yield, field spatial correlation of this magnitude was found in only three of the 10 data-sets and all of these related to mapping populations. Thus, the use of composite field samples may be a reasonable strategy for early generation milling trials, although more data need

to be considered in order to make this a general recommendation.

An important factor to consider in making the choice between testing individual field replicates or using composite samples is whether further processing, i.e. more experimental phases, are planned. In the case of wheat quality, traits associated with dough properties (such as dough strength and extension) are very important. These traits are obtained using a three-phase experiment in which the first two phases are as for flour yield, then in the third phase dough is made from the flour and tested for a range of traits. In the present authors' experience, the non-genetic variation in dough strength and extension traits has a large component due to field variation and spatial correlation (see Mann *et al*. 2006). Genetic gain for these traits may therefore be best served by maintaining individual field replicates through all phases of the experiment. Experience with the analysis of malting quality in barley has shown that the non-genetic variation in many of the traits is often dominated by field variation and spatial correlation (see Cullis *et al*. 2003) so that the testing of individual field replicates may be recommended. However, more experience with these traits is required in order to make firm recommendations.

Another interesting possibility is the use of phase confounded designs. This may be appropriate if a specific phase contributes very little to the total error variation in a trait. In this case, extra replication in the subsequent phase would be avoided so that, in terms of the analysis the experiment would then be regarded as having one less phase. An example of this type is the wheat dough property traits, the majority of which exhibit little or no variation from the milling phase (Mann *et al*. 2006). Thus, the milling and dough testing phases may be confounded with the result that the experiment is regarded as comprising two rather than three phases.

One key aspect that was not investigated in the simulation study was the ability to detect outliers (associated with any of the phases), which is only possible with some level of replication. In the detailed analysis of one of the data-sets, the detection and subsequent deletion of three (laboratory) outliers resulted in a large reduction (approximately 30 %) in effective error variance that would translate to a large increase in genetic gain. The approach to the detection of outliers presented in the present paper is informal, based mainly on graphical displays. Possible outliers are identified, then advice sought as to their likely cause and an appropriate remedy. A more objective approach to outlier detection is required. This is a difficult problem in the framework of linear mixed models and in particular for multi-phase data in which outliers may arise at several levels.

Another issue is the choice of values for p and q. In the simulation study, $p=q=0{\cdot}10$ but other values may be more appropriate. At present, it is recommended to use 'sensible' values that allow a statistical analysis to be conducted but do not lead to an excessive total number of samples. The choice of optimum values for individual traits is the subject of current research. The major unresolved issue, however, is that of optimal randomizations for p/q–rep designs. The literature on designs for multi-phase experiments is scarce and is confined to balanced, orthogonal cases. The issues for plant breeding data are even more complex. For example, the designs use partial rather than complete replication and there are complications induced by the fact that only a subset of the genotypes grown in the field trial is then tested in the laboratory (and the subset is unknown prior to designing the field trial). Currently, the present authors follow the lead of Wood *et al*. (1988) who suggest that a 'good' design is needed for each phase. Here 'good' refers to an optimality criterion on the genotypes. In the present paper, the decision was made to minimize the average pairwise prediction error variance of the genotypes (also see Cullis *et al*., in press), since this is equivalent to maximizing expected genetic gain. At present, this can only be done by assuming the field design is given, then ignoring field information in construction of the laboratory design. This is unlikely to be the optimum strategy. Note that the superiority of the naive p/q–rep design used in the simulation study would be even greater in terms of a more optimum design. The investigation of randomizations for multi-phase quality experiments is the subject of current research.

Finally, it is recognized that variety trials are usually conducted as series of experiments known as multi-environment trials (MET). Literature on the analysis of single-phase field MET data is expansive (see Smith *et al*. 2005 for a recent review) but is limited to only one or two papers in the context of multi-phase data (Cullis *et al*. 2003, for example). In the present paper, a mixed model analysis for a single multi-phase variety trial has been described. A future paper will describe an approach for multi-phase MET data that combines the multi-phase aspects presented here with the MET approach of Smith *et al*. (2001*b*). In their mixed model analysis (of single-phase data), Smith *et al*. (2001*b*) use a multiplicative model for variety by environment interaction and a separate spatial model for the (field) errors for each trial. This can be generalized to accommodate the modelling of both field and laboratory variation for each trial as described in the present paper. In terms of experimental design for multi-phase MET, note that the p/q–rep designs presented herein for individual trials are well suited to series of trials. The key principle is to seek balance across trials for the total number of samples for each variety. Thus varieties that are tested without replication in one trial would be replicated in

another (also see Cullis *et al*., in press, in the context of partially replicated field designs). The issue of optimal randomizations for multi-phase MET is a difficult problem and is the subject of current research.

We gratefully acknowledge the financial support of the Grains Research and Development Corporation of Australia (GRDC) and the NSW Centre for Agricultural Genomics. We thank Dr Peter Martin (NSW Department of Primary Industries) and the GRDC and Australian Winter Cereals Molecular Marker Programme for providing the example datasets. We also thank the participating Australian cereal chemistry laboratories. We thank referees for comments that have greatly improved the manuscript.

REFERENCES

Bainbridge, T. (1965). Staggered, nested designs for estimating variance components. *Industrial Quality Control* **22**, 12–20.

Becker, R. A., Chambers, J. M. & Wilks, A. R. (1988). *The New S Language*. Pacific Grove, CA, USA: Wadsworth and Brooks/Cole.

Brien, C. J. (1983). Analysis of variance tables based on experimental structure. *Biometrics* **39**, 53–59.

Brien, C. J. & Bailey, R. A. (in press). Multiple randomizations. *Journal of the Royal Statistical Society, Series B* **68**, 1–29.

Butler, D. G., Cullis, B. R., Gilmour, A. R. & Gogel, B. J. (2003). *Samm Reference Manual*. Training series, No QE02001. Brisbane, QLD, Australia: QLD Department of Primary Industries and Fisheries.

Coombes, N. E. (2002). *The reactive tabu search for efficient correlated experimental designs*. Ph.D. thesis, Liverpool John Moores University, Liverpool, U.K.

Cullis, B. R. & Gleeson, A. C. (1991). Spatial analysis of field experiments – an extension to two dimensions. *Biometrics* **47**, 1449–1460.

Cullis, B., Smith, A., Panozzo, J. & Lim, P. (2003). Barley malting quality: are we selecting the best? *Australian Journal of Agricultural Research* **54**, 1261–1275.

Cullis, B. R., Smith, A. B. & Coombes, N. E. (in press). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological and Environmental Statistics*.

Eckermann, P. J., Verbyla, A. P., Cullis, B. R. & Thompson, R. (2001). The analysis of quantitative traits in wheat mapping populations. *Australian Journal of Agricultural Research* **52**, 1195–1206.

Gilmour, A. R., Cullis, B. R. & Verbyla, A. P. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* **2**, 269–293.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., Welham, S. J. & Thompson, R. (2002). *ASReml User Guide*. Release 1.0. Hemel Hempstead, UK: VSN International Ltd.

Kempton, R. A. (1984). The design and analysis of unreplicated field trials. *Vortrage fur Pflanzenzuchtung* **7**, 219–242.

Kenward, M. G. & Roger, J. H. (1997). Small sample inference for fixed effects estimates from restricted maximum likelihood. *Biometrics* **53**, 983–997.

Mann, G., Allen, H., Morell, M. K., Nath, Z., Martin, P., Oliver, J., Cullis, B. & Smith, A. (2006). Comparison of small scale and large scale extensibility of dough produced from wheat flour. *Australian Journal of Agricultural Research* **56**, 1387–1394.

McIntyre, G. A. (1955). Design and analysis of two-phase experiments. *Biometrics* **11**, 324–334.

Nelder, J. A. (1965). The analysis of randomized experiments with orthogonal block structure. I. Block structure and the null analysis of variance. *Proceedings of the Royal Society, A* **283**, 147–162.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **55**, 545–555.

Smith, A. B., Cullis, B. R., Appels, R., Campbell, A. W., Cornish, G. B., Martin, D. & Allen, H. M. (2001*a*). The statistical analysis of quality traits in plant improvement programs with application to the mapping of milling yield in wheat. *Australian Journal of Agricultural Research* **52**, 1207–1219.

Smith, A. B., Cullis, B. R. & Thompson, R. (2001*b*). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* **57**, 1138–1147.

Smith, A. B., Cullis, B. R. & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science, Cambridge* **143**, 449–462.

Stram, D. O. & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* **50**, 1171–1177.

Verbyla, A. P., Eckermann, P. J., Thompson, R. & Cullis, B. R. (2003). The analysis of quantitative trait loci in multi-environment trials using a multiplicative mixed model. *Australian Journal of Agricultural Research* **54**, 1395–1408.

Wilk, M. B. & Gnanadeskan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55**, 1–17.

Wood, J. T., Williams, E. R. & Speed, T. P. (1988). Non-orthogonal block structure in two-phase designs. *Australian Journal of Statistics* **30**A, 225–237.