A Simulation Study on Tests of Hypotheses and Confidence Intervals for Fixed Effects in Mixed Models for Blocked Experiments With Missing Data

Joachim SPILKE, Hans-Peter PIEPHO, and Xiyuan HU

This article considers the analysis of experiments with missing data from various experimental designs frequently used in agricultural research (randomized complete blocks, split plots, strip plots). We investigate the small sample properties of REML-based Wald-type F tests using linear mixed models. Several methods for approximating the denominator degrees of freedom are employed, all of which are available with the MIXED procedure of the SAS System (8.02). The simulation results show that the Kenward-Roger method provides the best control of the Type I error rate and is not inferior to other methods in terms of power.

Key Words: Kenward-Roger method; Missing at random; Restricted maximum likelihood (REML); Satterthwaite method; Wald test.

1. INTRODUCTION

Data collected in agricultural experiments and surveys can often be considered as a realization $\mathbf{y}$ of a normally distributed random vector $\underline{\mathbf{y}}$, which follows a mixed linear model. Although the main focus usually is on inference for fixed effects, a realistic model frequently requires adding random effects to the linear predictor. In general form, the mixed linear model can be written as:

$$\underline{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\underline{\mathbf{u}} + \underline{\mathbf{e}} \tag{1.1}$$

(Henderson 1990, p. 1ff), where

$$\begin{aligned} \boldsymbol{\beta} &= p \times 1 \quad \text{vector of fixed effects} \\ \underline{\mathbf{u}} &= q \times 1 \quad \text{vector of random effects} \end{aligned}$$

Joachim Spilke is Associate Professor of Biometrics and Informatics in Agriculture, Agricultural Faculty, Martin-Luther-University Halle-Wittenberg, 06099 Halle, Germany (E-mail: spilke@landw.uni-halle.de). Hans-Peter Piepho is Associate Professor of Bioinformatics, Faculty of Agricultural Sciences, University of Hohenheim, 70599 Stuttgart, Germany (E-mail: piepho@uni-hohenheim.de). Xiyuan Hu is Associate Professor of Crop Science, Agricultural Faculty, University of Yangling, 712100 Yangling, China (E-mail:xiyuanhu@yahoo.com.cn).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 10, Number 3, Pages 374–389
DOI: 10.1198/108571105X58199

$$\begin{array}{rcll} \underline{\mathbf{e}} & = & n \times 1 & \text{vector of residuals} \\ \mathbf{X} & = & n \times p & \text{design matrix for fixed effects} \\ \mathbf{Z} & = & n \times q & \text{design matrix for random effects.} \end{array}$$

(Throughout this article, all random variables in model equations are underscored.) It is further assumed that

$$\begin{array}{rcl} \underline{\mathbf{u}} & \sim & N(0, \mathbf{G}), \\ \underline{\mathbf{e}} & \sim & N(0, \mathbf{R}), \\ E(\underline{\mathbf{y}}) & = & \mathbf{X}\boldsymbol{\beta},\ \text{var}(\underline{\mathbf{y}}) = \mathbf{ZGZ}' + \mathbf{R} = \mathbf{V}, \end{array}$$

and

$$\text{var}\begin{pmatrix} \underline{\mathbf{y}} \\ \underline{\mathbf{u}} \\ \underline{\mathbf{e}} \end{pmatrix} = \begin{bmatrix} \mathbf{V} & \mathbf{ZG} & \mathbf{R} \\ \mathbf{GZ}' & \mathbf{G} & 0 \\ \mathbf{R} & 0 & \mathbf{R} \end{bmatrix}.$$

Providing $\mathbf{G}$ and $\mathbf{R}$, and hence $\mathbf{V}$ are known, the best linear unbiased estimators (BLUE) of estimable functions $\mathbf{h}'\boldsymbol{\beta}$ of the fixed effects in (1.1) are given by

$$\mathbf{h}'\hat{\boldsymbol{\beta}} = \mathbf{h}' \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{1.2}$$

with

$$\text{var}(\mathbf{h}'\hat{\boldsymbol{\beta}}) = \mathbf{h}' \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-} \mathbf{h}. \tag{1.3}$$

When $\mathbf{V}$ is estimated from the data, estimators based on (1.2) are not generally BLUE (Henderson 1963) and exact tests of linear hypotheses are not usually available (Henderson 1984, p. 83), except in some balanced data settings. This problem is of great practical relevance, because in most applications $\mathbf{V}$ is unknown, and more often than not one is faced with unbalanced data.

Various procedures have been proposed for testing hypotheses on fixed effects in mixed models with unknown $\mathbf{V}$, most of which assume that $\mathbf{V}$ is estimated by the REML method (Giesbrecht and Burns 1985; Fai and Cornelius 1996; Kenward and Roger 1997). We compare several such methods by simulation for data structures characterized by missing data and small sample sizes.

2. IMPLICATIONS OF UNKNOWN G AND R

In most practical applications, $\mathbf{G}$ and $\mathbf{R}$ will be unknown. The implications for estimates of fixed effects based on (1.2) are as follows:

(1) The variance-covariance matrix $\mathbf{V}$ needs to be replaced by an estimate. Here, we restrict attention to REML estimation. The resulting estimates of fixed effects are often referred to as empirical BLUE (eBLUE). eBLUE do not usually have the BLUE properties, except in some balanced data settings. Specifically, they are no

longer linear functions of the observed data vector $\mathbf{y}$ and do not have the minimum variance property, though unbiasedness is still guaranteed, even for unbalanced data (Henderson 1963; Kackar and Harville 1981; 1984).

(2) The variance of an estimate of an estimable function may be a function of several components of variance, and tests of hypotheses require an approximation of the degrees of freedom.

(3) For unbalanced data, standard error estimates based on (1.3) with $\mathbf{V}$ replaced by its estimate are biased downwards (Henderson 1984; Kackar and Harville 1984). This problem is particularly relevant in small datasets, when many fixed effects need to be estimated.

With unbalanced data, ANOVA estimators of variance components lack optimality (minimum variance among all unbiased quadratic estimators), though unbiasedness is still guaranteed (Ahrens 1967; Searle 1971). Several alternative estimation methods have been proposed, for example, ML/REML (Hartley and Rao 1967; Patterson and Thompson 1971), and MINQUE/MIVQUE (Rao 1971; Lamotte 1973). Searle, Casella, and McCulloch (1992, p. 254) advocated use of ML or REML because of their near-optimal properties under normality of the data: consistency, asymptotic normality of estimators, and availability of asymptotic standard errors. The latter fact is exploited in approaches for approximating the degrees of freedom (Fai and Cornelius 1996; Kenward and Roger 1997). An advantage of REML over ML is the agreement with ANOVA estimators of variance components in balanced data, provided none of the ANOVA estimates is negative (Searle et al. 1992). This article uses REML throughout, because with unbalanced data, the ANOVA approach (Type III; Searle 1987, p. 391) leads to an inferior control of the Type I error rate in Wald-type F tests of fixed effects hypotheses based on eBLUE (Spilke und Tuchscherer 2001; Guiard, Spilke, and Dänicke 2003).

We have studied the performance of Wald-type tests based on eBLUE, with REML variance estimation of variance components. The REML-based method may be contrasted to an ANOVA approach to testing fixed effects in linear mixed models, which uses ordinary least squares estimation (OLSE). In some settings, the ANOVA-approach may perform quite well (Remmenga and Johnson 1995; Khuri, Mathew, and Sinha 1998), though due to the optimality of BLUE we expect eBLUE based on REML to be more efficient that OLSE in most cases. A thorough comparison by simulation would be rewarding, but is beyond the scope of the present article.

3. MATERIAL AND METHODS

3.1 INVESTIGATED EXPERIMENTAL DESIGNS

Because of their practical relevance, the following two-factorial experimental designs were considered.

Randomized complete block design:

$$\underline{y}_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \underline{bl}_k + \underline{e}_{ijk}. \tag{3.1}$$

Split-plot design, with main plots arranged in complete blocks:

$$\underline{y}_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \underline{bl}_k + \underline{f}_{ik} + \underline{e}_{ijk}. \tag{3.2}$$

Strip-plot design:

$$\underline{y}_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \underline{bl}_k + \underline{f}_{ik} + \underline{g}_{jk} + \underline{e}_{ijk}, \tag{3.3}$$

where $(i = 1, \ldots, a; j = 1, \ldots, b; k = 1, \ldots, r)$ with

$$\underline{bl} \sim N(0, \sigma^2_{bl});\ \underline{f} \sim N(0, \sigma^2_{RA});\ \underline{g} \sim N(0, \sigma^2_{RB});\ \underline{e} \sim N(0, \sigma^2_{R}),$$

where

μ	$=$	general mean
α_i	$=$	fixed effect of ith level of factor A
β_j	$=$	fixed effect of jth level of factor B
$(\alpha\beta)_{ij}$	$=$	fixed interaction of ith level of factor A and jth level of factor B
$\underline{bl}_k$	$=$	random block effect
$\underline{f}_{ik}$ and $\underline{g}_{jk}$	$=$	random main plot errors for rows and columns, respectively, within a block
$\underline{e}_{ijk}$	$=$	random residual (sub-plot) error

Independence of all random effects is assumed throughout. Here we are concerned with the case, where a balanced design becomes unbalanced by virtue of data missing at random. Such data will be classification-unbalanced (cell sizes in the $A \times B$ classification are not all equal) as well as variance-unbalanced (variances of pairwise treatment contrasts are not constant).

3.2 SIMULATED DATA STRUCTURES

For each of the three designs, we investigated three sizes $(a, b, r) : (2, 3, 3), (2, 3, 4)$, and $(3, 4, 4)$. Unbalancedness was generated by randomly dropping two observations. The variance ratios reflect results from 60 trials with winter barley and winter wheat. Ratios of variance components to residual error ranged from .1 to .3. In order to cover a broad spectrum of practically relevant settings, variance components σ^2_{bl}, σ^2_{RA}, and σ^2_{RB} were varied with values equal to .1, .5, 1, and 5 while the residual variance σ^2_R was fixed at unity. We studied four methods for approximating the degrees of freedom (see Section 3.3) under both the null and the alternative hypotheses. Thus, in total we simulated 288 different cases (3 designs $\times$ 3 sizes $\times$ 4 variance ratios $\times$ 4 approximation methods $\times$ 2 hypotheses).

For each case studied, we simulated 100,000 datasets. This simulation sample size will guarantee that the width of a 95% confidence interval for the empirical Type I error

Table 1. Cell and Marginal Means in Case of Validity of the Alternative Hypothesis

Size: a = 2, b = 3

Levels of factor A	*Levels of factor B* 1	2	3	*Marginal mean*
1	−1.5	−.5	.5	−.5
2	.5	.5	.5	.5
Marginal mean	−.5	0	.5	

Size: a = 3, b = 4

Levels of factor A	*Levels of factor B* 1	2	3	4	*Marginal mean*
1	−1.5	−1.0	0	.5	−.5
2	−.5	−.25	.25	.5	0
3	.5	.5	.5	.5	.5
Marginal mean	−.5	−.25	.25	.5	

rate is smaller than 5% of the nominal Type I error rate α, when $\alpha = .05$. In case of a valid alternative hypothesis, the interval width for the power may be somewhat larger depending on the true Type II error rate. Values of fixed effect used under the alternative are summarized in Table 1. All simulations and analyses were performed using the DATA step and the MIXED procedure of the SAS System (version 8.02).

3.3 APPROXIMATIONS OF THE DENOMINATOR DEGREES OF FREEDOM

The mixed model analysis for models (3.1)–(3.3) uses REML for estimating variance components and different methods for approximating the degrees of freedom. Fixed effects are estimated based on (1.2), with $\mathbf{V}$ replaced by a plug-in REML estimate. Null hypotheses of the form $H_0 : \mathbf{h}'\boldsymbol{\beta} = 0$ are tested by

$$t = \frac{\mathbf{h}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{h}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{h}}}, \tag{3.4}$$

when $\text{rank}(\mathbf{h}) = 1$ and by

$$F = \frac{\hat{\boldsymbol{\beta}}'\mathbf{h}\left(\mathbf{h}'\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-}\mathbf{h}\right)^{-1}\mathbf{h}'\hat{\boldsymbol{\beta}}}{\text{rank}(\mathbf{h})}, \tag{3.5}$$

when $\text{rank}(\mathbf{h}) > 1$. In general, the test statistics in (3.4) and (3.5) only have approximate t and F distributions, respectively. The approximate degrees of freedom ν for $t(\nu)$ and $F[\text{rank}(\mathbf{h}), \nu]$ were determined using four different methods as implemented in the MIXED procedure of SAS:

1. Residual method (SAS PROC MIXED option DDFM = residual): $\nu = n - \text{rank}(\mathbf{X})$.
2. Containment method (DDFM = contain): ν equals the "rank contribution" to $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$ of random effects which contain the fixed effect involved in the hypothesis.

For example, when a model has fixed effect A and random effects B and $A * B$, then $A * B$ contains A (SAS 1999).This is the default in MIXED.

3. Extended Satterthwaite (1941) method of Giesbrecht and Burns (1985) and Fai and Cornelius (1996) (DDFM = satterth).
4. Kenward-Roger method (Kenward and Roger 1997) (DDFM = kenwardroger): This approximation also uses the basic idea of Satterthwaite (1941). Its extension relative to the Satterthwaite method of Giesbrecht and Burns (1985) and Fai and Cornelius (1996) is an asymptotic correction of the estimated variance-covariance matrix of the fixed effects $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}$ due to Kackar and Harville (1984; see also Kenward and Roger 1997).

We considered these four methods because they are readily available in the MIXED procedure, and an informed choice is required on the part of the user.

3.4 INVESTIGATED CONTRASTS

The following contrasts of rank one were tested:

$$\text{C1} : \mu_{1.} - \mu_{2.} \quad = \quad \alpha_1 - \alpha_2 + \frac{1}{b}\sum_{j=1}^{b}\left[(\alpha\beta)_{1j} - (\alpha\beta)_{2j}\right]$$

(factor A main effects);

$$\text{C2} : \mu_{.1} - \mu_{.2} \quad = \quad \beta_1 - \beta_2 + \frac{1}{a}\sum_{i=1}^{a}\left[(\alpha\beta)_{i1} - (\alpha\beta)_{i2}\right]$$

(factor B main effects);

$$\text{C3} : \mu_{11} - \mu_{12} \quad = \quad \beta_1 - \beta_2 + (\alpha\beta)_{11} - (\alpha\beta)_{12}$$

($A * B$ means at same level of A);

$$\text{C4} : \mu_{11} - \mu_{21} \quad = \quad \alpha_1 - \alpha_2 + (\alpha\beta)_{11} - (\alpha\beta)_{21}$$

($A * B$ means at same level of B); and

$$\text{C5} : \mu_{11} - \mu_{22} \quad = \quad \alpha_1 - \alpha_2 + \beta_1 - \beta_2 + (\alpha\beta)_{11} - (\alpha\beta)_{22}$$

($A * B$ means at different levels of A and B).

Contrasts of rank greater than one were:

$$\text{C6} : \begin{bmatrix} \mu_{1.} - \mu_{2.} \\ \mu_{1.} - \mu_{3.} \\ \mu_{2.} - \mu_{3.} \end{bmatrix} = \begin{bmatrix} \alpha_1 - \alpha_2 + \frac{1}{b}\sum_{j=1}^{b}\left[(\alpha\beta)_{1j} - (\alpha\beta)_{2j}\right] \\ \alpha_1 - \alpha_3 + \frac{1}{b}\sum_{j=1}^{b}\left[(\alpha\beta)_{1j} - (\alpha\beta)_{3j}\right] \\ \alpha_2 - \alpha_3 + \frac{1}{b}\sum_{j=1}^{b}\left[(\alpha\beta)_{2j} - (\alpha\beta)_{3j}\right] \end{bmatrix}$$

(size $3 \times 4 \times 4$ only; factor A main effects);

$$\text{C7}: \begin{bmatrix} \mu_{.1} - \mu_{.2} \\ \mu_{.1} - \mu_{.3} \\ \mu_{.2} - \mu_{.3} \end{bmatrix} = \begin{bmatrix} \beta_1 - \beta_2 + \frac{1}{a}\sum_{i=1}^{a}\left[(\alpha\beta)_{i1} - (\alpha\beta)_{i2}\right] \\ \beta_1 - \beta_3 + \frac{1}{a}\sum_{i=1}^{a}\left[(\alpha\beta)_{i1} - (\alpha\beta)_{i3}\right] \\ \beta_2 - \beta_3 + \frac{1}{a}\sum_{i=1}^{a}\left[(\alpha\beta)_{i2} - (\alpha\beta)_{i3}\right] \end{bmatrix}$$

(level 1–3 factor B means);

$$\text{C8}: \begin{bmatrix} \mu_{11} - \mu_{12} \\ \mu_{11} - \mu_{13} \\ \mu_{12} - \mu_{13} \end{bmatrix} = \begin{bmatrix} \beta_1 - \beta_2 + (\alpha\beta)_{11} - (\alpha\beta)_{12} \\ \beta_1 - \beta_3 + (\alpha\beta)_{11} - (\alpha\beta)_{13} \\ \beta_2 - \beta_3 + (\alpha\beta)_{12} - (\alpha\beta)_{13} \end{bmatrix}$$

($A * B$ means at fixed level 1 of A);

$$\text{C9}: \begin{bmatrix} \mu_{11} - \mu_{21} \\ \mu_{11} - \mu_{31} \\ \mu_{21} - \mu_{31} \end{bmatrix} = \begin{bmatrix} \alpha_1 - \alpha_2 + (\alpha\beta)_{11} - (\alpha\beta)_{21} \\ \alpha_1 - \alpha_3 + (\alpha\beta)_{11} - (\alpha\beta)_{31} \\ \alpha_2 - \alpha_3 + (\alpha\beta)_{21} - (\alpha\beta)_{31} \end{bmatrix}$$

(size $3 \times 4 \times 4$ only; $A * B$ means at fixed level 1 of B); and

$$\text{C10}: \begin{bmatrix} \mu_{11} - \mu_{22} \\ \mu_{11} - \mu_{23} \end{bmatrix} = \begin{bmatrix} \alpha_1 - \alpha_2 + \beta_1 - \beta_2 + (\alpha\beta)_{11} - (\alpha\beta)_{22} \\ \alpha_1 - \alpha_2 + \beta_1 - \beta_3 + (\alpha\beta)_{11} - (\alpha\beta)_{23} \end{bmatrix}$$

($A * B$ means at different levels of A and B).

The contrasts were specified in the CONTRAST statement of MIXED using the exact specifications given above. For example, contrast C6 is specified as follows: CONTRAST "C6" A 1 –1 0, A 1 0 –1, A 0 1 –1. It should be noted that in mixed models this does not usually yield the same result as the statement CONTRAST "C6" A 1 –1 0, A 1 0 –1, when the Satterthwaite or the Kenward-Roger method is used. Results are the same, however, when using the containment and the residual methods. The contrasts C7, C8, C10 were studied for all design sizes ($2 \times 3 \times 4$ and $3 \times 4 \times 4$). The contrasts C6 and C9 were considered only for $3 \times 4 \times 4$ designs.

4. RESULTS

4.1 RESULTS UNDER THE NULL HYPOTHESIS

Due to space limitations, we present only part of the total simulation results for 288 settings. The choice will be justified in the relevant passages. The complete results are available from the first author upon request. Table 2 summarizes results for confidence intervals (coverage probabilities and widths) for all designs, sizes, and methods of approximating the df for contrasts of rank one. Note that the complement to the empirical coverage probability is equivalent to the Type I error rate, when data are simulated under the null hypothesis. Differences in interval width were mainly governed by the variance ratios. Short widths were usually associated with a large residual variances (ratio .1:1 etc.), while wider inter-

Table 2. Minima and Maxima of the Empirical Coverage Probability (CP) as Well as the Interval Width (IW) Across all Investigated Variance Ratios for *t*-test Under the Null Hypothesis (nominal confidence level = .95)

		$C1(\mu_{1.} - \mu_{2.})$		$C2(\mu_{.1} - \mu_{.2})$		$C3(\mu_{11} - \mu_{12})$		$C4(\mu_{11} - \mu_{21})$		$C5(\mu_{11} - \mu_{22})$	
Design	*Method*	*CL*	*IW*	*CL*	*IW*	*CL*	*IW*	*CL*	*IW*	*CL*	*IW*
						Size: 2 × 3 × 3					
Block	Residual	.937–.944	1.083–1.125	.937–.945	1.323–1.373	.939–.944	1.866–1.936	.937–.944	1.866–1.936	.937–.944	1.867–1.937
	Containment	.944–.951	1.121–1.164	.944–.951	1.369–1.421	.943–.950	1.932–2.004	.944–.951	1.931–2.003	.944–.951	1.932–2.004
	Satterthwaite	.940–.950	1.099–1.161	.941–.951	1.343–1.418	.940–.950	1.896–1.999	.941–.948	1.895–1.998	.940–.951	1.869–1.999
	Kenward-Roger	.942–.951	1.111–1.167	.943–.951	1.356–1.424	.942–.951	1.912–2.007	.942–.951	1.912–2.006	.942–.951	1.913–2.008
Split-plot	Residual	.842–.938	1.246–3.572	.900–.932	1.278–1.381	.921–.932	1.381–1.865	.884–.940	1.944–3.572	.884–.939	1.945–3.997
	Containment	.894–.972	1.553–4.451	.942–.951	1.403–1.517	.940–.951	1.979–2.137	.907–.957	2.135–4.389	.906–.956	2.136–4.390
	Satterthwaite	.927–.953	1.620–6.077	.933–.949	1.330–1.502	.932–.949	1.877–2.117	.924–.947	2.024–5.507	.924–.946	2.025–5.507
	Kenward-Roger	.928–.955	1.648–6.098	.937–.951	1.355–1.521	.935–.950	1.910–2.142	.940–.950	2.055–5.528	.925–.947	2.056–5.530
Strip-plot	Residual	.837–.908	1.191–3.577	.896–.930	1.462–3.983	.911–.915	1.865–4.048	.873–.896	1.781–4.048	.919–.935	2.033–5.660
	Containment	.889–.949	1.484–4.457	.944–.959	1.841–5.017	.989–.993	3.534–8.144	.974–.985	3.375–7.657	.992–.996	3.852–10.72
	Satterthwaite	.931–.947	1.669–6.176	.939–.951	1.676–4.910	.932–.942	1.979–4.975	.927–.939	2.015–5.788	.943–.946	2.141–6.498
	Kenward-Roger	.935–.952	1.726–6.273	.942–.955	1.721–4.981	.937–.946	2.036–5.071	.933–.945	2.082–5.905	.946–.950	2.194–6.580
						Size: 2 × 3 × 4					
Block	Residual	.941–.948	.883–.907	.945–.948	1.080–1.109	.941–.948	1.525–1.566	.940–.947	1.525–1.566	.940–.946	1.525–1.567
	Containment	.944–.952	.900–.925	.946–.952	1.100–1.130	.945–.951	1.554–1.596	.944–.951	1.554–1.596	.944–.950	1.555–1.597
	Satterthwaite	.943–.952	.891–.924	.945–.952	1.090–1.130	.944–.951	1.540–1.595	.943–.951	1.540–1.595	.942–.950	1.541–1.596
	Kenward-Roger	.944–.952	.896–.926	.945–.952	1.096–1.132	.944–.951	1.547–1.598	.944–.951	1.547–1.598	.943–.950	1.548–1.598
Split-plot	Residual	.871–.940	1.013–3.092	.931–.940	1.056–1.118	.932–.941	1.491–1.578	.898–.943	1.593–3.398	.899–.937	1.059–3.398
	Containment	.905–.967	1.169–3.569	.943–.951	1.110–1.175	.949–.951	1.567–1.658	.910–.953	1.657–3.571	.910–.953	1.675–3.571
	Satterthwaite	.938–.954	1.193–4.366	.938–.951	1.082–1.172	.939–.951	1.529–1.654	.932–.948	1.613–4.229	.933–.948	1.631–4.228
	Kenward-Roger	.939–.955	1.203–4.371	.940–.957	1.092–1.178	.941–.958	1.542–1.662	.933–.958	1.643–4.235	.934–.958	1.644–4.235
Strip-plot	Residual	.871–.925	.986–3.112	.915–.939	1.198–3.375	.925–.931	1.553–3.598	.898–.920	1.505–3.441	.930–.944	1.671–4.782
	Containment	.905–.953	1.138–3.592	.945–.965	1.386–3.592	.971–.977	2.026–4.693	.952–.970	1.963–4.489	.975–.983	2.180–6.239
	Satterthwaite	.939–.952	1.213–4.427	.943–.953	1.307–3.872	.936–.944	1.606–3.962	.934–.944	1.606–4.337	.945–.950	1.720–5.211
	Kenward-Roger	.942–.954	1.232–4.450	.946–.955	1.325–3.893	.940–.948	1.631–3.994	.937–.948	1.634–4.374	.948–.952	1.744–5.237

Table 2. Continued

		Size: 3 × 4 × 4									
		$C1(\mu_{1.} - \mu_{2.})$		$C2(\mu_{.1} - \mu_{.2})$		$C3(\mu_{11} - \mu_{12})$		$C4(\mu_{11} - \mu_{21})$		$C5(\mu_{11} - \mu_{22})$	
Design	*Method*	*CL*	*IW*	*CL*	*IW*	*CL*	*IW*	*CL*	*IW*	*CL*	*IW*
Block	Residual	.947–.951	.729–.735	.946–.950	.842–.848	.948–.952	1.457–1.468	.946–.950	1.457–1.468	.946–.950	1.457–1.468
	Containment	.947–.952	.732–.737	.947–.951	.845–.851	.948–.952	1.463–1.474	.947–.951	1.462–1.473	.947–.951	1.462–1.473
	Satterthwaite	.947–.952	.731–.737	.947–.951	.844–.851	.948–.952	1.461–1.474	.947–.951	1.461–1.473	.946–.952	1.460–1.473
	Kenward-Roger	.947–.952	.731–.737	.947–.951	.845–.851	.949–.952	1.462–1.474	.947–.951	1.462–1.473	.947–.951	1.462–1.473
Split-plot	Residual	.909–.940	.859–3.130	.943–.949	.833–.851	.943–.949	1.441–1.473	.923–.949	1.522–3.404	.924–.948	1.522–3.403
	Containment	.933–.961	.957–3.484	.946–.952	.844–.863	.946–.952	1.461–1.493	.926–.952	1.543–3.449	.927–.951	1.542–3.449
	Satterthwaite	.945–.954	.956–3.719	.945–.952	.839–.863	.945–.952	1.453–1.493	.944–.950	1.533–3.814	.945–.950	1.533–3.814
	Kenward-Roger	.945–.954	.957–3.720	.945–.952	.842–.863	.945–.952	1.457–1.494	.944–.951	1.536–3.815	.945–.950	1.536–3.815
Strip-plot	Residual	.907–.937	.856–3.129	.924–.943	.965–3.223	.932–.943	1.506–3.463	.922–.942	1.496–3.411	.938–.949	1.590–4.677
	Containment	.932–.960	.953–3.483	.947–.964	1.074–3.588	.941–.953	1.571–3.612	.932–.951	1.560–3.558	.947–.957	1.659–4.879
	Satterthwaite	.944–.954	.960–3.722	.946–.954	1.034–3.572	.943–.949	1.523–3.713	.942–.949	1.520–3.826	.947–.951	1.605–4.909
	Kenward-Roger	.945–.955	.963–3.723	.946–.955	1.037–3.573	.944–.950	1.529–3.715	.943–.949	1.526–3.828	.947–.951	1.610–4.910

Table 3. Minima and Maxima of the Empirical Type I Errors Across all Investigated Variance Cases for *F*-Test Under the Null Hypothesis (nominal Type I error = .05)

	Block			Split-plot			Strip-plot		
Contrast	*Containment*	*Satterthwaite*	*Kenward-Roger*	*Containment*	*Satterthwaite*	*Kenward-Roger*	*Containment*	*Satterthwaite*	*Kenward-Roger*
				Size: 2 × 3 × 3					
C7	.050–.055	.057–.063	.065–.071	.050–.057	.065–.079	.072–.085	.003–.012	.059–.078	.065–.089
C8	.050–.055	.057–.063	.065–.071	.052–.058	.066–.079	.073–.086	.004–.007	.077–.081	.080–.085
C10	.051–.056	.052–.060	.051–.059	.049–.085	.054–.061	.053–.064	.003–.005	.051–.058	.050–.053
				Size: 3 × 4 × 4					
C6	.050–.052	.050–.052	.054–.056	.061–.104	.052–.068	.056–.075	.056–.096	.052–.069	.057–.077
C7	.051–.052	.050–.052	.054–.056	.050–.054	.053–.057	.055–.059	.045–.067	.053–.070	.055–.068
C8	.049–.051	.048–.050	.053–.055	.049–.053	.051–.055	.059–.054	.046–.061	.055–.063	.059–.067
C9	.051–.052	.050–.052	.054–.056	.049–.082	.051–.066	.055–.072	.049–.075	.057–.069	.061–.074
C10	.051–.052	.051–.053	.051–.052	.052–.069	.051–.053	.052–.054	.043–.055	.050–.061	.051–.052

vals occured when the residual variance was small (variance ratio 5:1). Thus, as expected, large nonresidual variances caused wider intervals.

Although the residual method yielded the smallest width throughout, it also displayed the largest departure from the nominal confidence level. With the block design, all methods except the residual method were associated with satisfactory empirical coverage probabilities. More pronounced differences were observed for the split-plot design, where the containment method performed poorly with the contrasts $\mu_{1.} - \mu_{2.}$, $\mu_{11} - \mu_{21}$, and $\mu_{11} - \mu_{22}$, yielding too small coverage probabilities (excessive Type I error rates).

Similar results were found for the strip-plot design, where the nominal confidence level was exceeded also for the containment method, the sizes $2 \times 3 \times 3$ and $2 \times 3 \times 4$, and the contrast $\mu_{11} - \mu_{12}$. As expected, control of the nominal error probability improved with increasing sample size.

The Satterthwaite and Kenward-Roger methods provided the best control of the nominal coverage and error probabilities. The residual method performed poorly and will not be considered in the rest of this article. Despite unsatisfactory error control, the containment method will be considered further, because it is the default method of the MIXED procedure. Furthermore, we restrict our attention to the sizes $2 \times 3 \times 3$, which was the most unfavorable case, and $3 \times 4 \times 4$, the most favorable case.

Results for tests of hypotheses of rank greater than one shown in Table 3 indicate that the error control depends strongly on the design and the contrast. Although the nominal level was controlled well with block designs, considerable departures were observed with the containment method for the strip-plot design particularly for the size $2 \times 3 \times 3$. In addition, for the Satterthwaite and Kenward-Roger methods clear deviations were observed for certain contrasts (C7, C8) and this size. Generally, results for the different contrasts depended to a considerable degree on the variance ratio, as can be seen from the range of empirical coverage probabilities.

4.2 RESULTS UNDER THE ALTERNATIVE HYPOTHESIS

In order to facilitate interpretation of empirical power results, we report the noncentrality parameter for each contrast, computed by $\boldsymbol{\beta}'\mathbf{h}(\mathbf{h}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{h})^{-1}\mathbf{h}'\boldsymbol{\beta}$ for each simulation run and average across runs. Tables 4 and 5 report the minima and maxima of the noncentrality parameter across variance ratios. It should be stressed that results for the same contrast are not comparable between structures $2 \times 3 \times 3$ and $3 \times 4 \times 4$ due to differences in noncentrality. Also, it is noted that a relatively small residual variance (variance ratio 5:1 etc.) generally yields smaller power than relatively larger residual variances (variance ratio .1:1 etc.).

For rank-one hypotheses the empirical power follows roughly the same pattern as the empirical coverage probability under the null hypothesis (Table 2 and 4). When the empirical coverage probability was below the nominal level under the null hypothesis, the power was relatively high. A typical example is given by contrast $\mu_{1.} - \mu_{2.}$ for the split-plot design $2 \times 3 \times 3$ and the containment method. With a variance ratio of 5:5:1,

Table 4. Minima and Maxima of the Empirical Power Across all Investigated Variance Cases for *t*-test Under the Alternative Hypothesis as Well as the Noncentrality Parameter (NCP) (calculated by $\beta'h(h'(X'V^{-1}X)^{-}h)^{-1}h'\beta$)

	Block				Split-plot				Strip-plot			
	NCP	*Containment*	*Satterthwaite*	*Kenward-Roger*	*NCP*	*Containment*	*Satterthwaite*	*Kenward-Roger*	*NCP*	*Containment*	*Satterthwaite*	*Kenward-Roger*
						Size: 2 × 3× 3						
C1($\mu_{1.} - \mu_{2.}$)	3.66–3.75	.396–.421	.397–.435	.394–.428	.28–3.00	.132–.233	.083–.282	.082–.276	.27–2.94	.138–.274	.085–.296	.081–.255
C2($\mu_{.1} - \mu_{.2}$)	.62–.63	.109–.117	.110–.124	.109–.120	.60–.63	.103–.117	.106–.133	.103–.127	.07–.53	.061–.068	.066–.096	.063–.090
C3($\mu_{11} - \mu_{12}$)	1.27–1.29	.172–.183	.172–.191	.171–.188	1.23–1.28	.158–.180	.161–.202	.158–.195	.23–1.17	.018–.032	.081–.190	.076–.180
C4($\mu_{11} - \mu_{21}$)	5.08–5.16	.509–.534	.510–.548	.508–.542	.96–4.71	.196–.457	.155–.499	.153–.490	.93–4.67	.062–.175	.137–.504	.129–.485
C5($\mu_{11} - \mu_{22}$)	5.08–5.16	.507–.533	.509–.547	.506–.541	.96–4.71	.197–.459	.156–.499	.153–.489	.52–4.32	.019–.093	.103–.457	.099–.442
						Size: 3 × 4 × 4						
C1($\mu_{1.} - \mu_{2.}$)	1.88–1.89	.267–.272	.265–.270	.265–.270	.09–1.37	.075–.166	.075–.177	.059–.176	.09–1.36	.077–.171	.062–.177	.062–.175
C2($\mu_{.1} - \mu_{.2}$)	.35–.36	.090–.092	.089–.091	.089–.091	.35–.36	.089–.094	.089–.096	.089–.095	.02–.27	.055–.064	.056–.073	.056–.073
C3($\mu_{11} - \mu_{12}$)	.47–.48	.104–.106	.102–.105	.203–.106	.46–.47	.100–.106	.100–.107	.100–.107	.08–.43	.068–.096	.063–.106	.063–.105
C4($\mu_{11} - \mu_{21}$)	1.90–1.91	.269–.273	.266–.272	.267–.271	.36–1.73	.111–.245	.111–.249	.087–.248	.33–1.73	.104–.241	.087–.257	.087–.254
C5($\mu_{11} - \mu_{22}$)	2.96–2.97	.385–.391	.383–.389	.382–.388	.51–2.70	.134–.353	.134–.358	.106–.356	.28–2.48	.083–.312	.082–.335	.082–.333

Table 5. Minima and Maxima of the Empirical Power Across all Investigated Variance Cases for *F*-test Under the Alternative Hypothesis as Well as the Noncentrality Parameter (NCP) (calculated by $\boldsymbol{\beta}'\mathbf{h}(\mathbf{h}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{h}')^{-1}\mathbf{h}'\boldsymbol{\beta}$)

	Block				*Split-plot*				*Strip-plot*			
	NCP	*Containment*	*Satterthwaite*	*Kenward-Roger*	*NCP*	*Containment*	*Satterthwaite*	*Kenward-Roger*	*NCP*	*Containment*	*Satterthwaite*	*Kenward-Roger*
						Size: 2 × 3 × 3						
C7	2.49–2.54	.203–.220	.221–.238	.242–.263	2.40–2.51	.181–.209	.215–.261	.234–.278	.26–2.12	.014–.015	.092–.183	.103–.197
C8	5.14–5.20	.374–.400	.392–.418	.426–.459	5.00–5.17	.329–.374	.378–.442	.405–.472	.92–4.71	.014–.045	.135–.419	.144–.438
C10	6.83–6.93	.478–.507	.480–.528	.477–.522	1.01–6.16	.168–.405	.120–.450	.128–.443	.61–5.66	.009–.036	.083–.404	.081–.394
						Size: 3 × 4 × 4						
C6	7.54–7.55	.643–.650	.630–.637	.654–.661	.38–5.47	.136–.495	.089–.426	.099–.451	.38–5.45	.126–.466	.092–.427	.102–.449
C7	4.97–4.98	.459–.466	.449–.456	.471–.478	4.92–4.96	.445–.466	.447–.467	.464–.486	.47–4.16	.101–.362	.094–.385	.101–.400
C8	4.44–4.45	.418–.425	.406–.412	.430–.436	4.40–4.44	.404–.422	.403–.421	.422–.441	.77–4.04	.113–.362	.115–.395	.122–.413
C9	7.60–7.62	.646–.651	.624–.631	.656–.663	1.32–6.94	.193–.600	.157–.592	.169–.620	1.31–6.92	.176–.583	.159–.607	.170–.631
C10	6.18–6.19	.550–.558	.550–.559	.550–.558	1.25–5.54	.172–.503	.134–.506	.141–.505	.55–5.08	.094–.434	.086–.469	.087–.467

the coverage probability under the null hypothesis was .894, which is clearly below the nominal (Table 2), and, accordingly, the power (.132) was considerably larger compared to the Satterthwaite and Kenward-Roger methods (.083 and .082, respectively; Table 4). For the same example with a variance ratio of .1:.1:1, the Type I error probability was markedly exceeded. Accordingly, the power (.233) was high compared to the Satterthwaite and Kenward-Roger methods (.282 and .276, respectively). The same general pattern was also observed for hypotheses of rank greater than one. For example, the empirical Type I error rate was below the nominal level for all contrasts with the strip-plot design $2 \times 3 \times 3$ and the containment method (Table 3). Thus, the higher power with the Kenward-Roger method is mainly a result of the higher Type I error rates.

5. CONCLUSION

In simulations, we considered three experimental designs, looking at different contrasts and different variance ratios. Overall, the Kenward-Roger method yielded the smallest range and the smallest bias of empirical Type I error rates and was not inferior in terms of power. Thus, this method can be recommended.

Tests of linear hypotheses $\mathbf{h}'\boldsymbol{\beta} = \mathbf{0}$ in the mixed model are not generally exact when fixed effects are estimated by generalized least squares with estimated $\mathbf{V}$. Also, the plug-in estimate of the standard error, $\sqrt{[\mathbf{h}'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{h}]}$, will be biased downwards. Although full results are not reported here for brevity, we should emphasize that the degree of bias depended on data structure, sample size, degree of imbalance, ratio of variance components, experimental design, and the contrast of interest. The bias in the standard error estimates was considerably reduced by using the Kackar-Harville correction as described by Kenward and Roger (1997). The magnitude of the correction was positively correlated with the degree of bias, though the bias is not entirely removed.

Analysis by the containment method yielded satisfactory control of the Type I error rate only in special cases such as the randomized complete block design, where the only random effect—that is, that for blocks—was not contained in the treatment effects. When at least one of the random effects contained a treatment effect, confidence levels and thus Type I errors may be severely biased. By contrast, the Satterthwaite method provided good control of the Type I error rate. For tests of rank one hypotheses (t-tests), the Kenward-Roger method reduced the bias in the estimated variance-covariance matrix of linear contrasts, which is the main reason it gave the best control of the Type I error rate. For tests of hypotheses with rank greater than one, we observed a small advantage in favour of the Satterthwaite method.

The power analysis showed the same pattern as that of the Type I error rate, that is, power was larger when the Type I error rate was on the liberal side. Thus, differences in power among the methods were mainly due to differences in Type I error control. The main result of our article is that the Kenward-Roger method is competitive in terms of power, when all methods yield satisfactory Type I error control. The power differences between the contrasts can be explained mainly by differences in the noncentrality parameter.

Based on our simulation results, we recommend the Kenward-Roger method for the linear mixed model analysis of designed experiments with missing data. It should be stressed, however, that further simulations need to be performed for other designs and other settings, including more complex variance-covariance structures and multivariate data. Piepho (1997), considering unbalanced subsampling data from blocked experiments, found reasonable performance of the Satterthwaite method. Kenward and Roger (1997) found good performance of their method across a number of designs. These results are in good agreement with our findings. Keselman, Kowalchuk, Algina, and Wolfinger (1999) reported on a simulation with repeated-measures designs common in behavioral science research, in which the Satterthwaite method is compared with a Welch-James-type test. The authors' results did not generally favor one approach over the other. Schaalje, McBride, and Fellingham (2002) investigated repeated-measures designs with five covariance structures. In their simulation study, the Kenward-Roger method worked as well as or better than the Satterthwaite method in all situations and produced Type I error rates close to the nominal values in case of compound symmetry and Toeplitz structures. When the covariance structure became more complex (first-order-antedependence), even the Kenward-Roger method had problems and produced inflated error rates. Thus, the Kenward-Roger method should be used with particular caution, when the random part of the model does not have a simple random effects structure as the models considered in the present article.

ACKNOWLEDGMENTS

We appreciate the referees comments, which led to a considerable number of improvements. Furthermore, we thank the Deutsche Forschungsgemeinschaft for funding the third author.

[Received July 2003. Revised September 2003.]

REFERENCES

Ahrens, H. (1967), *Die Varianzanalyse*, Berlin: Akademie-Verlag.

Fai, A. H. T., and Cornelius, P. L. (1996), "Approximate F-Tests of Multiple Degree of Freedom Hypotheses in Generalized Least Squares Analyses of Unbalanced Split-Plot Experiments," *Journal of Statistical Computing and Simulation*, 54, 363–378.

Giesbrecht, F. G., and Burns, J. C. (1985), "Two-Stage Analysis Based on a Mixed Model: Large-Sample Asymptotic Theory and Small-Sample Simulation Results," *Biometrics*, 41, 477–486.

Guiard, V., Spilke, J., and Dänicke, S. (2003), "Evaluation and Interpretation of the Results for Three Cross-Over Designs," *Archives of Animal Nutrition*, 57, 177–195.

Hartley, H. O., and Rao, C. R. (1967), "Maximum Likelihood Estimation for the Mixed Analysis of Variance Model," *Biometrika*, 54, 93–108.

Henderson, C. R. (1963), "Selection Index and Expected Genetic Advance," *Statistical Genetics and Plant Breeding*, NAS-NRC Publ. No.982, 141–163.

——— (1984), *Application of Linear Models in Animal Breeding*, Guelph: University of Guelph.

——— (1990), "Statistical Method in Animal Improvement: Historical Overview," in *Advances in Statistical Methods for Genetic Improvement of Livestock*, New York: Springer.

Kackar, A. N., and Harville, D. A. (1981), "Unbiasedness of Two-Stage Estimation and Precision Procedures for Mixed Linear Models," *Communications in Statistics A*, 10, 1249–1261.

——— (1984), "Approximation for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models," *Journal of the American Statistical Association*, 79, 853–861.

Kenward, M. G., and Roger, J. H. (1997), "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood," *Biometrics*, 53, 983–997.

Keselman, H. J., Kowalchuk, R. K., Algina, J., and Wolfinger, R. D. (1999), "The Analysis of Repeated Measurements: A Comparison of Mixed-Model Satterthwaite F tests and a Nonpooled Adjusted Degrees of Freedom Multivariate Test," *Communications in Statistics A*, 28, 2967–2999.

Khuri, A. I., Mathew, T., and Sinha, B. K. (1998), *Statistical Tests for Mixed Linear Models*, New York: Wiley.

Lamotte, L. R. (1973), "Quadratic Estimation of Variance Components," *Biometrics*, 29, 311–330.

Patterson, H. D., and Thompson, R. (1971), "Recovery of Inter-Block Information When Block Sizes are Unequal," *Biometrika*, 58, 545–554.

Piepho, H. P. (1997), "Analysis of a Randomized Complete Block Design with Unequal Subclass Numbers," *Agronomy Journal*, 89, 718–723.

Rao, C. R. (1971), "Minimum Variance Quadratic Unbiased Estimation of Variance Components," *Journal of Multivariate Analysis*, 1, 445–456.

Remmenga, M. D., and Johnson, D. E. (1995), "A Comparison of Inference Procedures in Unbalanced Split-Plot Designs," *Journal of Statistical Computation and Simulation*, 51, 353–367.

SAS Institute Inc. (1999), SAS OnlineDoc®, Version 8, Cary, NC: SAS Institute Inc.

Satterthwaite, F. E. (1941), "Synthesis of Variance," *Psychometrika*, 6, 309–316.

Schaalje, G. B., McBride, J. B., and Fellingham, G. W. (2002), "Adequacy of Approximation to Distributions of Test Statistics in Complex Mixed Linear Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 512–524.

Searle, S. R. (1971), "Topics in Variance Component Estimation," *Biometrics*, 27, 1–76.

——— (1987), *Linear Models for Unbalanced Data*, New York: Wiley.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: Wiley.

Spilke, J., and Tuchscherer, A. (2001), "Simulationsuntersuchungen zum Einfluss verschiedener Strategien der Varianzkomponentenschätzung und Hypothesenprüfung auf die statistischen Risiken in gemischten linearen Modellen mit ungleicher Klassenbesetzung," *Zeitschrift für Agrarinformatik*, 4, 66–75.