Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 8

5-1-2007

Type I Error Rates of the Kenward-Roger Adjusted Degree of Freedom F-test for a Split-Plot Design with Missing Values

Miguel A. Padilla
University of Alabama at Birmingham

James Algina
University of Florida, algina@ufl.edu

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 66- 80

1538 – 9472/07/$95.00

Type I Error Rates of the Kenward-Roger Adjusted Degree of Freedom *F*-test for a Split-Plot Design with Missing Values

Miguel A. Padilla
University of Alabama at Birmingham

James Algina
University of Florida

The Type I error rate of the Kenward-Roger (KR) test, implemented by PROC MIXED in SAS, was assessed through a simulation study for a one between- and one within-subjects factor split-plot design with ignorable missing values and covariance heterogeneity. The KR test controlled the Type I error well under all of the simulation factors, with all estimated Type I error rates between .040 and .075. The best control was for testing the between-subjects main effect (error rates between .041 and .057) and the worst control was for the between-by-within interaction (.040 to .075). The simulated factors had very small effects on the Type I error rates, with simple effects in two-way tables no larger than .01.

Key words: Missing values, Kenward-Roger *F*-test, robustness, mixed models, split-plot design.

Introduction

According to Keselman et al. (1998), one of the most commonly used designs in educational and psychological research is the split-plot design, a design which includes both between-subjects and within-subjects factors. Responses on the within-subjects factor are obtained by repeatedly measuring each participant in the study. The repeated measures might be obtained at different points in time or under different treatments. Unfortunately, data collected in split-plot designs can be incomplete for a variety of reasons. Consider participants who drop out of a longitudinal study because of illness or death, refuse to answer questions on a survey because of its length or the sensitivity of the questions, or are unable to answer questions on a performance assessment test because of time constraints or lack of ability. Each results in missing values.

Miguel A. Padilla is a postdoctoral fellow in the School of Public at the University of Alabama at Birmingham. His research interests are in applied statistics. Email him at mpadilla@uab.edu. James Algina is a Professor of Educational Psychology at the University of Florida. His research interests are in applied statistics and psychometrics. Email him at algina@ufl.edu.

Little and Rubin (2002, p. 12) and Rubin (1976) defined three types of missing data mechanisms. The missing data mechanisms, ordered from most restrictive to least restrictive in terms of assumptions made about the process that leads to the missing data, are missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Generally, the NMAR missing data condition constitutes any missing data condition that is not MCAR or MAR. Let $f(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi})$ denote the distribution of the missing data indicators for participant *i*, where $\boldsymbol{r}_i$ is a $K \times 1$ vector whose elements are zero for missing and one for observed in the corresponding elements of the $K \times 1$ vector of repeated variables $\boldsymbol{y}_i$, $\boldsymbol{X}_i$ is the design matrix for the factors, and $\boldsymbol{\psi}$ contains the parameters for the relationship of $\boldsymbol{r}_i$ to $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$.

Data are MCAR if $f(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{\psi})$, that is, if the distribution of the missing data indicators does not depend on the repeated measures. The $\boldsymbol{y}_i$ vector can be partitioned as $\boldsymbol{y}_i = [\boldsymbol{y}_{io}' \quad \boldsymbol{y}_{im}']'$ where $\boldsymbol{y}_{io}$ contains the repeated measures variables on which participant *i* has observed scores and $\boldsymbol{y}_{im}$ contains the repeated measures variables on which participant *i* has missing scores. If $f(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}) = f(\boldsymbol{r}_i \mid \boldsymbol{y}_{io}, \boldsymbol{X}_i, \boldsymbol{\psi})$, that is, the missing data indicator does not depend on the variables of which participant *i*

has missing scores, then the data are MAR.

The distribution of $\boldsymbol{y}_i$ can be written as $f(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ contains the main effect and interaction parameters as well as the parameters for the covariance matrix for the repeated measures. A general method for consistent maximum likelihood (ML) estimation of $\boldsymbol{\theta}$ is obtained by including both the observed scores on the repeated measures and the missing data indicators, as well as $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, in the likelihood. However, Rubin (1976) showed that if the missing data mechanism is MCAR or MAR and if the parameters $\boldsymbol{\psi}$ and $\boldsymbol{\theta}$ are disjoint, ML estimators of the $\boldsymbol{\theta}$ parameters are consistent when the missing data indicators and $\boldsymbol{\psi}$ are excluded from the data analysis.

Excluding the missing data indicators and $\boldsymbol{\psi}$ is referred to as ignoring the missing data mechanism. Thus, the MCAR or MAR missing data mechanisms are ignorable for purposes of ML estimation. If the data are MCAR, both listwise deletion and ML ignoring the missing data mechanism will produce consistent estimators, but the ML estimators will be more accurate because they use all of the available data. Rubin (1976) also showed that the MCAR missing data mechanism is ignorable for sampling distribution based inference procedures such as hypothesis tests and confidence intervals. So, if the data are MCAR, either listwise deletion or ML ignoring the missing data mechanism can be used for inference, but ML will result in more powerful tests and narrower confidence intervals because it does not delete the observed data for participants with some missing values.

When ML estimation is used, whether the MAR missing data mechanism is ignorable for sampling distribution based inference depends on how the sampling covariance matrix is calculated. The MAR missing data mechanism is ignorable for sampling distribution based inferences on the means if the sampling covariance matrix is estimated from the observed information matrix for the means and the covariance parameter estimates, but not if the matrix is estimated from the portion of the observed information matrix that pertains only to the means (Kenward & Molenberghs, 1998).

The MAR mechanism may not be ignorable for sampling distribution based inferences if the sampling covariance matrix is estimated from the expected information matrix. If the expected information matrix is used, it must take into account the actual sampling process implied by the MAR mechanism (Kenward & Molenberghs, 1998). Kenward and Molenberghs (1998) referred to using this type of expected information matrix as using the unconditional sampling framework; whereas using the information matrix that ignores this sampling process is referred to as using the naïve sampling framework.

If the missing data mechanism is NMAR, the missing data mechanism is non-ignorable for purposes of ML estimation, and the pattern of missing values must be taken into account to obtain consistent ML estimates. This can be accomplished by using a selection model that incorporates a model for the missing values indicator or by using a pattern mixture model, which stratifies the data on the basis of the pattern of missing values (Albert & Follmann, 2000; Algina & Keselman, 2004a, 2004b; Diggle & Kenward, 1994; Fitzmaurice, Laird, & Shneyer, 2001; Kenward, 1998; Little, 1995; Troxel, 1998). Little (1995) provided details about these two approaches.

Unfortunately, traditional methods for analyzing data from a split-plot design such as ANOVA, adjusted degrees of freedom ANOVA, and MANOVA use listwise deletion and therefore are not likely to yield valid inferences except when the missing data mechanism is MCAR, an often unrealistic assumption in applied settings. Furthermore, these tests also assume that the covariance matrices ($\boldsymbol{\Sigma}_j, j = 1, \ldots J$) are homogenous across the J levels of the between-subjects factor, another often-unrealistic assumption. The tests will often fail to control the Type I error when the homogeneity assumption is violated (Keselman & Keselman, 1990; Keselman, Keselman, & Lix, 1995; Keselman, Lix, & Keselman, 1996). For further details about these tests, see Greenhouse-Geisser (1959), Huynh and Feldt (1976), Huynh and Feldt (1970), Keselman and Keselman (1993), Mendoza (1980), and Looney and Stanley (1989).

As a response to the unsatisfactory results created by violating the homogeneity of covariance assumption required by the standard *F*-tests, the multivariate Welch-James (WJ) test, which does not require the sphericity assumption or the homogeneity of covariance assumption, has been proposed for use in split-plot designs (Algina & Keselman, 1997, 1998; Keselman, Algina, Wilcox, & Kowalchuk, 2000; Keselman, Carriere, & Lix, 1993). The WJ test tends to control the Type I error rates for the within-subjects main effect and the between- by within-subjects interaction whether or not the dispersion matrices are heterogeneous. However, the WJ test also utilizes listwise deletion when there are missing values and would be expected to yield valid inferences only when the missing data are MCAR.

The Kenward-Roger (KR) adjusted degrees of freedom *F*-test is similar to the WJ test, but uses all available data in parameter estimation when there are missing values. Because parameter estimation is carried out by ML, the estimated parameters are consistent when data are MCAR or MAR. Additionally, the KR test is computed through a mixed-effects linear model so multisample sphericity is not required and heterogeneity of covariance can be modeled. Furthermore, the KR test uses a more accurate estimator of the sampling covariance matrix than the standard mixed model *F*-test.

When the mixed-effects linear model is used to analyze data, likelihood ratio, score, or Wald hypothesis tests can be used. Wald tests seem to be the most common. For example, when PROC MIXED in SAS is used, the default procedure for tests on the fixed effects is the Wald test. Let $\mathbf{L}$ be a $r \times JK$ contrast matrix of full row rank and let $\boldsymbol{\mu} = [\boldsymbol{\mu}_1' \quad \boldsymbol{\mu}_2' \quad \cdots \quad \boldsymbol{\mu}_J']'$. Each $\boldsymbol{\mu}_j$ is a $K \times 1$ vector of population means for the K levels of the within-subjects factor in the split-plot design. The main effect and interaction hypotheses about the between- and within-subjects factors can be expressed as

$$H_0: \mathbf{L}\boldsymbol{\mu} = \mathbf{0} \tag{1}$$

where $\mathbf{0}$ is a $r \times 1$ vector with all elements equal to zero. Let $\boldsymbol{\Sigma}_j$ denote the $K \times K$ population covariance matrix of the repeated measures for the j^{th} level of the between-subjects factor, $\mathbf{S}_j$ the $K \times K$ restricted ML (REML) estimate of the covariance matrix and $\boldsymbol{\Sigma}_{ij}$ and $\mathbf{S}_{ij}$ the $K_i \times K_i$ sections ($i = 1, 2, \ldots, n_j$) of the population and sample covariance matrices, respectively that pertain to the dependent variables on which the i^{th} participant in the j^{th} group has observed scores. In addition let $\mathbf{A}_{\text{i}}$ denote a $K_i \times K$ indicator matrix obtained by eliminating the k^{th} ($k = 1, 2, \ldots, K$) row from the $K \times K$ identity matrix if the data for the i^{th} participant is missing on the k^{th} level of the within-subjects factor. The PROC MIXED default test statistic for testing the null hypothesis is

$$F = \frac{\bar{\mathbf{y}}'\mathbf{L}'\left(\mathbf{L}\hat{\mathbf{M}}^{-1}\mathbf{L}'\right)^{-1}\mathbf{L}\bar{\mathbf{y}}}{r} \tag{2}$$

where $\bar{\mathbf{y}} = [\bar{\mathbf{y}}_1' \quad \bar{\mathbf{y}}_2' \quad \cdots \quad \bar{\mathbf{y}}_J']'$ is the ML estimate of $\boldsymbol{\mu}$, r = rank($\mathbf{L}$), and $\hat{\mathbf{M}}$ is a block diagonal matrix in which the j^{th} block is $\sum_i \mathbf{A}_i'\mathbf{S}_{ij}^{-1}\mathbf{A}_i$. The vector

$$\bar{\mathbf{y}}_j = \left(\sum_i \mathbf{A}_i'\mathbf{S}_{ij}^{-1}\mathbf{A}_i\right)^{-}\left(\sum_i \mathbf{A}_i'\mathbf{S}_{ij}^{-1}\mathbf{y}_i\right).$$

The matrix $\hat{\mathbf{M}}^{-1}$ is the estimated sampling covariance matrix of the mean vector $\bar{\mathbf{y}}$ and is based on the expected information matrix calculated under the naïve sampling framework. Even when data are MCAR or there are no missing data, using $\hat{\mathbf{M}}^{-1}$ has two drawbacks:

1. $\hat{\mathbf{M}}^{-1}$ tends to be too small because it fails to take into account the uncertainty in $\bar{\mathbf{y}}$ introduced by substituting $\mathbf{S}_{ij}$ for $\boldsymbol{\Sigma}_{ij}$ when $\bar{\mathbf{y}}$ is obtained (Kackar & Harville, 1984).

2. $\hat{\mathbf{M}}^{-1}$ is a biased estimate of $\mathbf{M}^{-1}$ (Prasad & Rao, 1990; Booth & Hobert 1998). Harville and Jeske (1992) developed a better estimator of $\mathbf{M}^{-1}$, denoted by $\hat{\boldsymbol{m}}^{@}$. Kenward and Roger (1997) then developed an alternative estimator of $\mathbf{M}^{-1}$, denoted by $\hat{\boldsymbol{\Phi}}_A$.

Kenward and Roger (1997) also developed the test statistic

$$F^* \sim \lambda \frac{\bar{\mathbf{y}}'\mathbf{L}'\left(\mathbf{L}\hat{\mathbf{\Phi}}_A\mathbf{L}'\right)^{-1}\mathbf{L}\bar{\mathbf{y}}}{r} \quad (3)$$

where λ is a scaling factor and $F_{r,d}$, is the critical value where d is the approximate degrees of freedom. Both λ and d are estimated from the data. The Kenward-Roger procedure is implemented in SAS's PROC MIXED, but uses $\hat{\boldsymbol{m}}^{@}$ in place of $\hat{\mathbf{\Phi}}_A$.

Keselman et al. (1993) and Algina and Keselman (1997) investigated the performance of the WJ test at controlling the Type I error rate in a split-plot design under several simulation conditions. In the former study the authors investigated (a) the number of levels of the within-subjects factor (K = 4, 8), (b) the ratio of total sample size N to $K - 1$ (i.e., $N/(K - 1)$), (c) the ratio of the smallest n_j to $K - 1$ (i.e., $n_{\min}/(K - 1)$), (d) sample size inequality, (e) pairing of n_j with covariance matrices, and (f) the shape of the distribution of the data. In all conditions the number of levels of the between-subjects factor was three (J = 3) and heterogeneity of covariance matrices was held constant at a ratio of 1:3:5.

The latter study added J = 6, degree of departure from sphericity measured by epsilon (ε), and heterogeneity of covariance matrices with a ratio of 1:5:9. The authors were interested in the sample sizes required to control the Type I error rate when testing the within-subjects main effect and the between- by within-subjects interaction. In the first study, the sample sizes ranged from 30 to 171 and in the second study they ranged from 20 to 714. From these two studies the authors provided sample size guidelines for the WJ test to control the Type I error under normal and non-normal data. The final sample size recommendations are summarized in Table 1.

Fai and Cornelius (1996) developed and compared four alternative test statistics $(F_1 \text{ to } F_4)$ that can be used to test linear hypotheses on means in multivariate studies. They showed how to use the data to estimate the denominator degrees of freedom for the four statistics and the scaling factors λ_2 and λ_4 for the F_2 and F_4 statistics. The F_1 and F_2 statistics use $\hat{\mathbf{M}}^{-1}$ to estimate the covariance matrix of the mean vector whereas F_3 and F_4 use $\hat{\boldsymbol{m}}^{@}$. The F_4 statistic is similar to the statistic obtained by using the KR option in PROC MIXED, but with a different formula for the scaling factor and the degrees of freedom. The F_1 test is available in SAS when the Satterthwaite option is used in PROC MIXED. For further details on F_1 through F_2 see Fai and Cornelius.

(3)

Fai and Cornelius (1996) applied their tests to split-plot designs with a three-level between-subjects factor (J) and a four-level within-subjects factor (K). The covariance structure was compound symmetric. The design was unbalanced in that the number of subjects varied across levels of the between-subjects factor and data were not generated for some combinations of subjects and the within-subjects factor. Because the missing data were never generated, the missing data mechanism was effectively MCAR. Estimated Type I error rates and power were reported for the main effect of the between-subjects factor. All four tests provided reasonable control of the Type I error rate. The performance of F_1 and F_3, which do not include a scaling factor were very similar. Type I error rates and power for F_4 was always larger than for F_3.

Schaalje, McBride, and Fellingham (2002), reporting on a study conducted by McBride (2002), reported Type I error rates for F_1 and the test obtained using the KR option in PROC MIXED. McBride investigated the performance of these tests in a split-plot design. The following provides a social science example of the design investigated by McBride. Suppose three methods for structuring interactions among students in a mathematics classroom are to be compared; n schools are randomly assigned to each method, where n was three in half of the conditions studied by McBride and five in the other half. The methods will be implemented for three, six, or nine weeks. Each school contributes K classes. Each class is assigned a single interaction quality score. In half of the conditions studied by McBride, K = 3 and the

Table 1. Final $n_{min}/(K-1)$ Recommendations for Distribution by Between-Subjects Factor (J) by Test by Within-Subjects Factor (K)

			$n_{min}/(K-1)$	
Distribution	J	Test	$K = 4$	$K = 8$
Normal	3	K	2.00	3.00
		$J \times K$	3.00	4.00
	6	K	1.33	1.43
		$J \times K$	4.75	5.00
Non-normal	3	K	3.00	4.00
		$J \times K$	8.00	6.00
	6	K	1.33	1.71
		$J \times K$	14.00	10.14

Note. Based on Keselman et al. (1993) and Algina and Keselman (1997)

design was balanced. In the other half, $K = 5$ so that within each school two classes would be assigned to two of the implementation periods and one class would be assigned to the remaining implementation period. In these conditions the design was unbalanced, but no data were missing.

McBride also investigated the effect of the covariance structure, which included the following five structures: compound symmetric (equal correlations and equal variance for the repeated measures), heterogeneous compound symmetric (equal correlations, but unequal variances for the repeated measures), Toeplitz, heterogeneous first-order autoregressive (correlations conform to a first-order autoregressive pattern, but the variances for the repeated measures are unequal), and first-order ante-dependence (see Wolfinger, 1996, for examples of these covariance structures). The results indicated that employing the KR option provided better control than did employing the Satterthwaite option in PROC MIXED. Type I error rates were closer to the nominal level for balanced designs than for unbalanced designs. For unbalanced designs, Type I error rates improved as n increased.

Kenward and Roger (1997) investigated how well the original KR procedure controlled Type I error rates in four situations: (a) a four-treatment, two-period cross-over design, (b) a row-column-α design, (c) a random coefficients regression model for repeated measures data, and (d) a split-plot design. In (c) and (d) there were missing data. In (c) the missing data mechanism was MCAR. The missing data mechanism in (d) was not specified. In all situations, the KR test controlled the Type I error rate well.

Kowalchuk, Keselman, Algina, and Wolfinger (2004) compared the performance of the KR and the WJ procedures at controlling the Type I error rate under several simulation conditions for a (J = 3) × (K = 4) split-plot design. The simulation conditions they

investigated were (a) type of population covariance structure, (b) degree of group size inequality, (c) positive and negative pairings of covariance matrices and group sample sizes, (d) shape of the data, and (e) type of covariance structure fit to the data. All simulation conditions had heterogeneous covariance matrices across the levels of the between-subjects factor (J) with a ratio of 1:3:5. Data with missing values were not investigated. The KR test coupled with modeling the true covariance structure of the data performed better than did the WJ test under all conditions with small sample sizes. Also, the authors showed that always assuming an unstructured covariance structure performed comparably to modeling the true covariance structure when using the KR test.

Based on the previous results, the KR test and similar tests like the F_4 test (Fai & Cornelius, 1996) can control the Type I error rate for a variety of repeated measures designs when there are either missing data but no covariance heterogeneity or covariance heterogeneity but no missing data. The purpose of this study is to investigate control of the Type I error rate by the KR test as it is implemented in PROC MIXED when there are both missing data and covariance heterogeneity. Because of the similarities between the KR test and the WJ test and because Type I error rates for the WJ test have been extensively evaluated by Algina and Keselman and their colleagues, the KR test will be evaluated under conditions similar to those used by these authors to evaluate the WJ test, with the addition of missing value conditions.

Methodology

Study Variables

Eight variables were manipulated in this simulation. The variables of interest are (a) the number of levels of the between-subjects factor (A), (b) the number of levels of the within-subjects factor (B), (c) $n_{\min}/(K-1)$ where K is the number of levels of the within-subjects factor, (d) sample size inequality across the between-subjects factor (SSI), (e) degree of sphericity as quantified with Box's (1954) epsilon (ε), (f) nature of pairing of group sizes with covariance matrices (NPSC), (g) type of missing data mechanism (TMDM), and (h) percent of missing data (PM). For each combination of levels of the factors, five thousand replications were generated.

Both the number of levels of the between-subjects and within-subjects factors were investigated in the study. Each of these factors had two levels with $J = 3, 6$ and $K = 3, 6$. In the initial planning, the study was going to investigate $J = 3, 6$ and $K = 4, 8$, but preliminary simulations indicated that using PROC MIXED took an inordinate amount of time when $K = 8$.

The sample sizes investigated were $n_{\min}/(K-1) = 4, 6$ for $J = 3$ and $n_{\min}/(K-1) = 5, 7.7$ for $J = 6$. Within each pair of $n_{\min}/(K-1)$ ratios, the smaller ratio corresponds to sample size recommendations in Table 1 for the between- by within-subjects interaction with normal data, $K = 8$, and $J = 3, 6$. The larger $n_{\min}/(K-1)$ values were based on the recommendations from Table 1 and the higher demands missing values will place on the data analysis.

Keselman et al. (1998) found that unequal sample sizes in split-plot designs were common, occurring in a little over 50% of the split-plot designs. For this reason unequal sample sizes were investigated. In particular, moderate and severe group size inequalities were investigated as defined by Keselman et al. (1993) through the coefficient of variation:

$$C = \left(\bar{n}\sqrt{J}\right)^{-1}\sqrt{\sum_{j=1}^{J}\left(n_j - \bar{n}\right)^2}, \quad (4)$$

where $C \simeq .16, .33$ describe moderate and severe group size inequality, respectively.

Departures from sphericity quantified by Box's (1954) epsilon (ε), were also investigated with $\varepsilon = .60, .75, .90$, where $\varepsilon = .60$ and $\varepsilon = .75$ represent relatively severe and moderate violations of sphericity, respectively. In past studies $\varepsilon = .40, .57, .75$ were investigated (Algina & Keselman, 1997; Keselman, Keselman, & Shaffer, 1991; Algina & Oshima, 1994). However, ε has a lower bound of $\varepsilon = 1/(K-1)$, so for $K = 3$ the lower bound is $\varepsilon = .50$ and so $\varepsilon = .40$ cannot be investigated. Also, according to Huynh and Feldt (1976) $\varepsilon =$

.75 represents the lower limit of ε found in educational and psychological data. The epsilon values in this simulation study were chosen based on this contention. In particular, note that $\varepsilon = .75$ is the mid value and the other values are $\varepsilon \pm .15$. The actual covariance matrices are shown in Table 2.

ratio of sample size to heterogeneity of covariance matrices was set at 1:3:5 for $J = 3$ and 1:3:5:1:3:5 for $J = 6$ (Algina & Keselman, 1997; Keselman et al., 1993; Keselman, Algina, Kowalchuk, & Wolfinger, 1999). Furthermore,

Table 2. Pooled Covariance Matrices

	$K = 3$	$K = 6$
$\varepsilon = .90$	$\begin{bmatrix} 18.0 & 5.0 & 6.0 \\ & 8.0 & 5.0 \\ & & 7.0 \end{bmatrix}$	$\begin{bmatrix} 18.0 & 5.0 & 7.0 & 7.0 & 6.0 & 5.0 \\ & 12.0 & 8.0 & 7.0 & 6.0 & 5.0 \\ & & 10.0 & 6.0 & 6.0 & 5.0 \\ & & & 10.0 & 5.0 & 5.0 \\ & & & & 9.0 & 5.0 \\ & & & & & 8.0 \end{bmatrix}$
$\varepsilon = .75$	$\begin{bmatrix} 23.2 & 4.5 & 7.4 \\ & 10.3 & 5.3 \\ & & 4.3 \end{bmatrix}$	$\begin{bmatrix} 29.6 & 12.7 & 7.5 & 7.0 & 5.9 & 5.9 \\ & 15.1 & 7.9 & 6.0 & 6.4 & 4.9 \\ & & 13.2 & 6.9 & 6.0 & 5.4 \\ & & & 9.4 & 6.0 & 4.8 \\ & & & & 8.0 & 5.0 \\ & & & & & 5.9 \end{bmatrix}$
$\varepsilon = .60$	$\begin{bmatrix} 23.8 & 1.9 & 9.3 \\ & 9.5 & 5.7 \\ & & 3.9 \end{bmatrix}$	$\begin{bmatrix} 28.8 & 4.8 & 10.1 & 9.8 & 8.3 & 7.3 \\ & 17.4 & 8.1 & 7.4 & 6.9 & 4.1 \\ & & 9.9 & 7.7 & 6.5 & 5.7 \\ & & & 8.3 & 5.6 & 4.3 \\ & & & & 5.6 & 4.4 \\ & & & & & 4.3 \end{bmatrix}$

The pairing direction, positive or negative, between the unequal group sizes and the heterogeneous covariance matrices were also investigated. A pairing is positive when the largest n_j is paired with the covariance matrix with the largest elements and negative when the largest n_j is paired with the covariance matrix with the smallest elements. In order to have comparability with previous research results, the previous studies have shown that this ratio and pairing can have a strong impact on the Type I error rate for approximate univariate *F*-tests, such as the Huynh-Feldt *F*-test (1976), and multivariate tests, particularly when the sample size is small (Keselman & Keselman, 1990). Specifically, positive pairings produce conservative Type I errors and negative pairing produce liberal Type I errors.

The MCAR and MAR missing data mechanisms were investigated in connection with 5%, and 15% probability of missing data at each level of the within-subjects factor except the first level; there were no missing data in the first level (see the Data Generation section for an explanation). Only the MCAR and MAR missing data mechanisms were investigated because Padilla and Algina (2004) demonstrated that the NMAR missing data mechanism negatively impacts the Type I error rate of the KR test statistic in a repeated measured design with no between-subjects factors.

Data Generation

The data were generated by using the model

$$y_{ijk} = \mu + e_{ijk} . \qquad (5)$$

The mean vector $\boldsymbol{\mu}_j = [\mu_1 \quad \mu_2 \quad \ldots \quad \mu_K]'$ was the same for all J groups and the elements μ_k were equal because the focus of the study was on control of the Type I error rate by the KR test. The common elements were arbitrarily set to zero. The **e** vector was a $K \times 1$ random vector such that $\mathbf{e} \sim NID(\mathbf{0}, \boldsymbol{\Sigma}_j)$.

All data simulations and analyses were conducted using SAS version 9.0. For each combination of levels of the simulation factors, the following steps were used to simulate the data in the j^{th} level of the between-subjects factor.

1. Simulate **Z**, a $n_j \times K$ matrix of pseudorandom standard normal variables where n_j is the sample size for the j^{th} level of the between-subjects split-plot design.

2. Calculate **T** a $K \times K$ upper triangular Cholesky factor of the covariance matrix $\boldsymbol{\Sigma}$.

3. Calculate $\boldsymbol{y} = d_j\mathbf{ZT}$, where d_j is a constant selected to create the required degree of covariance heterogeneity.

4. In all conditions there were no missing values on y_{i1}:

a. For MCAR, eliminate y_{ik} (k = 2, . . . , K) if $U_{ik} < \pi$ where π is the expected proportion of the missing data on y_k and U_{ik} is a uniform random variable.

b. For MAR, eliminate y_{ik} if $U_{ik} < \Phi(my_{i1} + c)$, where Φ is the cumulative standard normal distribution function and the parameters m and c will be described below.

In selecting data points for elimination, the parameter m controls how dependent the missing data are on y_1 in the MAR condition and was set to one. Let

$$r_{ik} = \begin{cases} 1 & \text{if } y_{ik} \text{ is missing} \\ 0 & \text{otherwise} \end{cases} .$$

With m = 1, the biserial correlation between r_{ik} and y_1 was .5 in the MAR condition. Hence, the missing data indicators depend fairly heavily on y_1. With m = 1, the expected proportion of missing data on y_k is dependent on c. In the procedure described in the preceding paragraphs, the probability that $r_{ik} = 1$ is related to y_1 is modeled by a normal ogive (probit) model. Using well-known facts about the normal ogive model (see, for example, Lord & Novick, 1968, equations 16.9.3 and 16.94), it can be shown that

$$c = \left\{\Phi^{-1}(\pi)\right\}\sqrt{1+m^2} . \qquad (6)$$

Thus, for m = 1, and for 5% and 15% missing data conditions, the expression becomes $c = -1.645\sqrt{2}$, and $c = -1.036\sqrt{2}$, respectively.

Data Analysis

The SAS PROC MIXED program used in this simulation is

```
proc mixed;
  class Person A B;
  model score = A B A*B/ ddfm=kenwardroger;
  repeated B/ subject=Person group=A type=un;
run;
```

The following list describes various aspects of the code.

- **Person** is a variable that identifies simulated subjects.

- **Score** is the variable containing scores on the dependent variable.
- **A** is a variable that identifies the levels of the between-subjects factor.
- **B** is a variable that identifies the levels of the within-subjects factor.
- **ddfm = kenwardroger** instructs SAS to use the KR statistic to test the main effects and the interaction.
- **Repeated** is a key word that tells SAS that B is a repeated measures (within-subjects) factor and is necessary when there are missing data.
- **Group = A** tells SAS to model the covariance matrix for each level of A. That is, it specifies modeling heterogeneity of covariance matrices across the levels of A.
- **Subject = Person** tells SAS that the score values are correlated within each person.
- **Type = un** instructs SAS to estimate an unstructured covariance matrix with K estimated variances and $K(K-1)/2$ estimated covariances.

Although there are several covariance structures that can be used to model the covariance matrix (Wolfinger, 1996), only the unstructured between-subjects heterogeneous structure (UN-H) covariance matrix was used in this simulation. Although using a UN-H covariance structure comes at the cost of estimating $K(K+1)/2$ parameters, Kowalchuk et al. (2004) showed that under similar simulation conditions assuming an unstructured covariance structure performed comparably to modeling the true covariance structure when using the KR test.

The corresponding *p*-values of applying the KR test to 5,000 replications were available for each combination of the investigated conditions. The result of each test was summarized by a dichotomous variable, defined in the following manner:

$$Type\ I\ Error = \begin{cases} 0 \text{ if the } p-\text{value} < .05 \\ 1 \text{ otherwise} \end{cases}.$$

For each of the between-subjects, within-subjects, and between- by within-subjects KR tests the Type I error variable was analyzed by using logistic regression with the study variables as factors. A forward selection approach was used to select appropriate models. The models used were an intercept-only model, a model with main effects only, a model with main effects and two-way interactions, and so forth. A model was considered adequate for the data if the χ^2 goodness of fit test was non-significant or if Bentler's (1990) Comparative Fit Index (*CFI*) ≥ .95. An index of fit was used because, due to the large number of replications, the χ^2 goodness of fit statistic for the logistic model could be very sensitive to small effects of the factors. The *CFI* in this context was calculated as follows:

$$CFI = 1 - \left(\lambda / \lambda_i\right) \qquad (7)$$

where $\lambda = \max(\chi^2 - df, 0)$, χ^2 and df are the chi-squared goodness of fit statistic for the fitted model and the corresponding degrees of freedom, $\lambda_i = \max(\chi_i^2 - df_i, \chi^2 - df, 0)$, and χ_i^2 and df_i is the chi-squared goodness of fit statistic for the intercept-only model and its corresponding degrees of freedom.

Assessment of the Type I error rates were based on Bradley's (1978) liberal criterion for identifying conditions in which hypothesis testing procedures work adequately. His liberal criterion is $.5\alpha \le \tau \le 1.5\alpha$ where α is the nominal Type I error and τ is the actual Type I error. Using α = .05, the liberal criterion is $.025 \le \tau \le .075$.

Results

Analysis of Type I Error Rates for the Between-Subjects Main Effect

The distribution of Type I error rates for the between-subjects main effect is shown in Figure 1 and has M = .050 and SD = .003. The range of the Type I error rate is [.041, .057]. The goodness of fit test for the intercept-only model was not significant, $\chi^2(383) = 398.64$, p = .28, suggesting that the effects of the factors were quite small. Because the Type I errors rates for the between-subjects main effect were predominately within Bradley's liberal criterion and because the intercept only model could not be rejected, it appears that the KR between-subjects omnibus test controls the Type

I error well at all levels of the investigated factors in this study.

The distribution of Type I error rates for the within-subjects main effect is shown in Figure 2 and has M = .052 and SD = .005. The range of the Type I error rate is [.041, .070]. Hence, in all conditions the Type I error rate was well within Bradley's (1978) liberal criterion interval. CFI for the model with main effects and two-way interactions was .98. In addition the goodness of fit test was non-significant, $\chi^2(339) = 354.24$, $p = .27$. Thus, the two-way interaction model was selected for further analysis. Wald tests indicated that all factors that had significant main effects also entered into significant two-way interactions. As might be expected from Figure 2, all effects were small. Mean Type I Error rates were between .048 and .061 in all two-way tables and no simple effect was as large as .01. Type I error rates tended to be larger when *J*, *K,* and percent missing data were larger. Type I error rates also tended to be larger for MAR data[1].

Analysis of Type I Error Rates for the Within-Subjects Main Effect

Because a major focus of this study is the effect of sample size on Type I error rates, two-way tables of means for the only interactions with sample size are presented in Table 3. These results indicate that control of the Type I error rate was good regardless of the sample size and that the effect of sample size on the Type I error rate was quite small.

Analysis of Type I Error Rates for the Between- by Within-Subjects Interaction

The distribution of Type I error rates for the interaction effect is presented in Figure 3 and has M = .054 and SD = .007. The range of the Type I error rate is [.040, .075]. Consequently, in all conditions the Type I error rate was once again within Bradley's liberal criterion interval. CFI for the model with main effects and two-way interactions was 1.00. In addition, the goodness of fit test was non-significant, $\chi^2(339) = 368.79$, $p = .23$. Wald tests indicated that all factors that had significant main effects also entered into significant two-way interactions. As might be expected from Figure 3, all effects of factor were small. Mean Type I Error rates were between.049 and .058 in all two-way tables and no simple effect was as large as .01. Type I error rates tended to be larger when *K,* sample size inequality, and percent missing were larger. Type I error rates also tended to be larger when the sample size-covariance pairing was negative. The effect of *J* was miniscule. The effect of type of missing data tended to be small and to vary in direction over levels of the factors with which it interacted.

Because a major focus of this study is the effect of sample size on Type I error rates, two-way tables of means for the interactions only with sample size are presented in Table 4. These results indicate that control of the Type I error rate was good regardless of the sample size and that the effect of sample size on the Type I error rate was quite small.

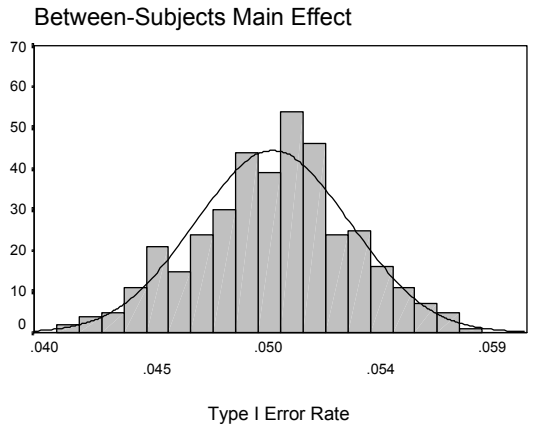

Figure 1. Distribution of Type I Error Rates: Between-Subjects KR *F*-Test

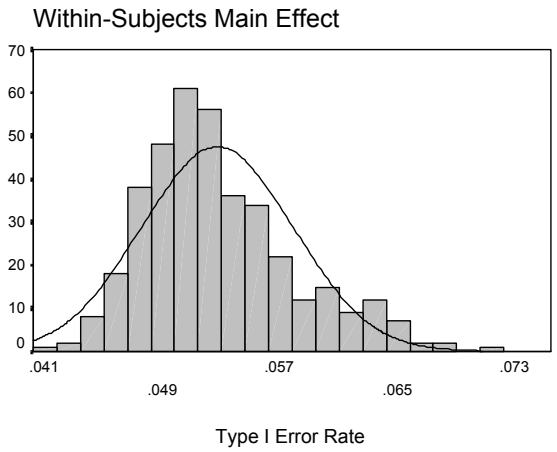

Figure 2. Distribution of Type I Error Rates: Within-Subjects KR *F*-Test

Table 3. Effect of $n_{min}/(K-1)$ on Type I Error Rates for the Within-Subjects Main Effect

		$n_{min}/(K-1)$	
Factor	Factor levels	Small	Large
K	3	.0503	.0514
	6	.0541	.0539
PM	5%	.0494	.0509
	15%	.0550	.0545

Note. Each proportion is out of 480,000 hypothesis tests.

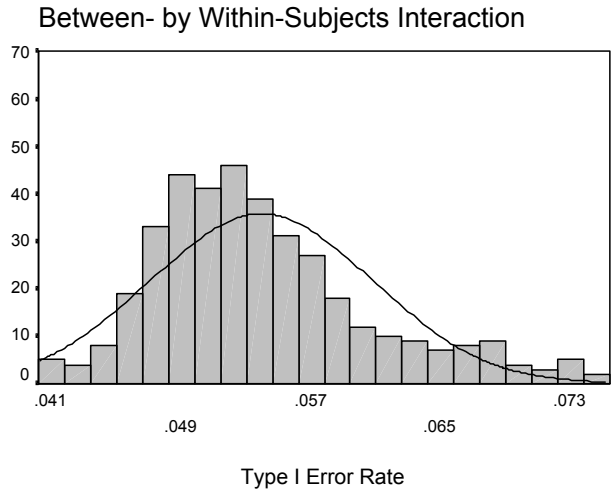

Figure 3. Distribution of Type I Error Rates: Interaction KR *F*-Test

Table 4. Effect of $n_{\min}/(K-1)$ on Type I Error Rates for the Between- by Within-Subjects Interaction

		$n_{min}/(K-1)$	
Factor	Factor levels	Small	Large
K	3	.0509	.0507
	6	.0577	.0553
SSI	.16	.0527	.0524
	.33	.0559	.0537
NPSC	Positive	.0509	.0517
	Negative	.0577	.0543
TMDM	MCAR	.0552	.0519
	MAR	.0534	.0542
PM	5%	.0508	.0504
	15%	.0578	.0557

Note. Each proportion is out of 480,000 hypothesis tests.

Conclusion

The results of this study support the conclusion that sampling distribution based inferences on the means using ML estimates can control the Type I error rate under MCAR missing data mechanisms. Additionally, sampling distribution based inferences using ML estimates can control the Type I error rate when the missing data mechanism is MAR (Little & Rubin, 2002; Rubin, 1976) Most important this control can be obtained with relatively modest sample size requirements.

With respect to the between-subjects main effect, the KR test statistic controlled the Type I error rate well across all levels of the simulation factors. Most Type I error rates were within both Bradley's conservative criterion and all were well within the liberal criterion. None of the simulation factors affected the Type I error rate of the between-subjects main effect. In regard to the within-subjects main effect and the within- by between-subjects interaction, although a number of factors affected Type I error rates, all effects were very small and all Type I error rates were within Bradley's liberal criterion.

The effects of the factors on Type I error rates were generally quite small. Nevertheless it is clear that the effects of the factors on the on Type I error rates must be due to their effects on the accuracy of the *F*-distribution as an approximation to the sampling distribution of the test statistic. The KR test statistic was selected because it uses a better estimator of the covariance matrix for small sample sizes and Satterthwaite (1946) type degrees of freedom based on the better estimate of the covariance matrix. However, when the data are incomplete in addition to being relatively small and paired with a MAR missing data mechanism, the accuracy of the approximation may be worse than when the data are complete.

Although the design investigated in this study was a popular split-plot design with one between- and one within-subjects factor, the positive findings open the door for further simulation work on using ML to directly estimate model parameters from split-plot designs with missing values. One condition that can be investigated is a non-normal distribution of the dependent variable. In the present study, the data were generated under a multivariate normal distribution and since data from educational or psychological research cannot be presumed to be normal, investigation of a non-normal data condition can provide applied researchers with valuable information as to whether the KR test is robust to the normality assumption. In other words, can the KR test control the Type I error when the normality assumption is violated?

Even though all of the Type I error rates of the KR test were within Bradley's (1978) liberal criterion, it is not clear at what percent of missing data the KR test will begin to breakdown. Additionally, it is not clear how small the sample sizes can be and still have the KR test provide reasonable control of the Type I error. Consequently, future work could focus on what are the percent of missing data and sample size requirements needed for the KR test to provide reasonable control of the Type I error.

An alternative to the estimator of the sampling covariance matrix used in the KR test is the sandwich estimator (White, 1980, Liang & Zeger, 1986). The sandwich estimator provides a consistent estimator of the covariance matrix given that the model for the means is correct. That is the model for the covariance structure need not be correct. Hence, it may be fruitful to compare the performance of the *F*-test using the sandwich estimator to the KR test at controlling the Type I error in a simulation study with ignorable missing data.

References

Albert, P. S., & Follmann, D. A. (2000). Modeling repeated count data subject to information dropout. *Biometrics, 56*, 667-677.

Algina, J., & Keselman, H. J. (1997). Testing repeated measure hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test. *Multivariate Behavioral Research, 32*, 255-274.

Algina, J., & Keselman, H. J. (1998). A power comparison of the Welch-James and Improved General Approximation tests in the split-plot design. *Journal of Educational and Behavioral Statistics, 23*, 152-169.

Algina, J., & Keselman, H. J. (2004a). A comparison of methods for longitudinal analysis with missing data. *Journal of Modern Applied Statistical Methods, 3*, 13-26.

Algina, J., & Keselman, H. J. (2004b). Assessing treatment effects in randomized longitudinal two-group designs with missing observations. *Journal of Modern Applied Statistical Methods, 3*, 271-287.

Algina, J., & Oshima, T. C. (1994). Type I error rates for Huynh's general approximation and improved general approximation tests. *British Journal of Mathematical and Statistical Psychology, 47*, 151-165.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Booth, J. G., & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association, 96*, 262-272.

Box, G. E. P. (1954). Some theorem on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance and correlations between errors in the two-way classification. *Annals of Mathematical Statistics, 25*, 484-498.

Bradley, J. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Diggle, P. D., & Kenward, M. G. (1994). Informative dropout in longitudinal data analysis. *Applied Statistics, 43*, 49-93.

Fai, H. T., & Cornelius, P. L. (1996). Approximate *F*-tests for multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation, 54*, 363-378.

Fitzmaurice, G. M., Laird, N. M., & Shneyer, L. (2001). An alternative parameterization of the general linear mixture model for longitudinal data with non-ignorable drop-outs. *Statistics in Medicine, 20*, 1009-1021.

Greenhouse, S. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Harville, D. A., & Jeske, D. R. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association, 87*, 724-731.

Huynh, H., & Feldt, L. S. (1970). Conditions under which mean square ratios in repeated measures designs have exact *F*-distributions. *Journal of the American Statistical Association, 65*, 1582-1589.

Huynh, H. & Feldt, L. S. (1976). Estimation of the Box correction for degree of freedom from sample data in randomized block and split-plot designs. *Journal of Educational Statistics, 1*, 69-82.

Kackar, R. N., & Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association, 79*, 853-862.

Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: An illustration of sensitivity. *Statistics in Medicine, 17*, 2723-2732.

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science, 13*, 236-247.

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

Keselman, H. J., Algina, J., Kowalchuk, R. K., & 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 – Theory and Methods, 28*, 2967-2999.

Keselman, H. J., Algina, J., Wilcox, R. R., & Kowalchuk (2000). Testing repeated measures hypothesis when covariance matrices are heterogeneous: Revisiting the robustness of the Welch-James test again. *Educational and Psychological Measurement, 60*, 925-938.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated measure hypotheses when covariance matrices are heterogeneous. *Journal of Educational Statistics, 18*, 305-319.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donohue, B., Kowalchuk, R. K., Lowman, L. L., Petosky, M. D., Keselman, J. C., & Levin, J. R. (1998).

Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68*, 350-386.

Keselman, H. J., & Keselman, J. C. (1990). Analysing unbalanced repeated measures designs. *British Journal of Mathematical and Statistical Psychology, 43*, 265-282.

Keselman, H. J., & Keselman, J. C. (1993). Analysis of repeated measurement. In L. K. Edwards (Eds.), *Applied analysis of variance in behavioral science* (pp. 105-145). New York: Marcel Dekker.

Keselman, H. J., Keselman, J. C., & Lix, L. M. (1995). The analysis of repeated measurements: Univariate tests, multivariate tests, or both? *British Journal of Mathematical and Statistical Psychology, 48*, 319-338.

Keselman, H. J., Keselman, J. C., & Shaffer, J. P. (1991). Multiple pairwise comparisons of repeated measures means under violation of multisample sphericity. *Psychological Bulletin, 110*, 162-170.

Keselman, J. C., Lix, L. M., & Keselman, H. J. (1996). The analysis of repeated measurements: A quantitative research synthesis. *British Journal of Mathematical and Statistical Psychology, 49*, 275-298.

Kowalchuk, R. K., Keselman, H. J., Algina, J., & Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted *F*-tests. *Educational and Psychological Measurement, 64*, 224-242.

Liang, K-Y., & Zeger, S. L. (1986). Longitudinal analysis using generalized linear models. *Biometrika, 73*, 13-22.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90*, 1112-1121.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2th ed.). New York: John Wiley & Sons.

Looney, S. W., & Stanley, W. B. (1989). Exploratory repeated measures analysis for two or more groups: Review and update. *American Statistician, 43*, 220-225.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test score, with contributions by Alan Birnbaum*. Reading, MA: Addison-Wesley.

McBride, G. B. (2002). Statistical methods helping and hindering environmental science and management. *Journal of Agricultural, Biological, and Environmental Statistics, 7*, 300-305.

Mendoza, J. L. (1980). A significance test for multisample sphericity. *Psychometrika, 45*, 495-498.

Padilla, M. A., & Algina, J. (2004). Type I error rates for a one factor within-subjects design with missing values. *Journal of Modern Applied Statistical Methods, 3*, 406-416.

Prasad, N. G. N, & Rao, J. N. K. (1990). The estimation of mean squared error of small-area estimators. *Journal of the American Statistical Association, 85*, 163-171.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581-592.

Satterthwaite, F. E. (1946). An approximation distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110-114.

Schaalje, G. B., McBride, J. B., & Fellingham, G. W. (2002). Adequacy of approximations to distributions of test statistics in complex mixed linear models. *Journal of Agricultural, Biological, and Environmental Statistics, 7*, 512-524.

Troxel, A. B. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics, 47*, 425-438.

White, H. (1980). A heterscedasticity-consistent covariance matrix and a direct test for heterscedasticity. *Econometrica, 48*, 817-838.

Wolfinger, R. D. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics, 1*, 205-230.