Journal of Modern Applied Statistical Methods

Volume 7 | Issue 2 Article 4

11-1-2008

Type I Error Rates of the Kenward-Roger *F*-test for a Split-Plot Design with Missing Values and Non-Normal Data

Miguel A. Padilla
Old Dominion University, mapadill@odu.edu

YoungKyoung Min
Korea Foundation for the Advancement of Science and Creativity, ykymin@yahoo.com

Guili Zhang
East Carolina University, zhangg@ecu.edu

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Padilla, Miguel A.; Min, YoungKyoung; and Zhang, Guili (2008) "Type I Error Rates of the Kenward-Roger *F*-test for a Split-Plot Design with Missing Values and Non-Normal Data," *Journal of Modern Applied Statistical Methods*: Vol. 7 : Iss. 2 , Article 4.
DOI: 10.22237/jmasm/1225512180

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 385-397
Copyright © 2008 JMASM, Inc.
1538 – 9472/08/$95.00

Type I Error Rates of the Kenward-Roger *F*-test for a Split-Plot Design with Missing Values and Non-Normal Data

Miguel A. Padilla
Old Dominion University

YoungKyoung Min
Korea Foundation for the Advancement of Science and Creativity

Guili Zhang
East Carolina University

The Type I error of the Kenward-Roger (KR) *F*-test was assessed through a simulation study for a between- by within-subjects split-plot design with non-normal ignorable missing data. The KR-test for the between- and within-subjects main effect was robust under all simulation variables investigated and when the data were missing completely at random (MCAR). This continued to hold for the between-subjects main effect when data were missing at random (MAR). For the interaction, the KR *F*-test performed fairly well at controlling Type I under MCAR and the simulation variables investigated. However, under MAR, the KR *F*-test for the interaction only provided acceptable Type I error when the within-subjects factor was set at 3 and 5% missing data.

Keywords: missing values, Kenward-Roger *F*-test, robustness, mixed models, split-plot design, non-normal data, and covariance heterogeneity.

Introduction

Linear mixed-effects, or mixed models, have become increasingly popular in analyzing data from split-plot designs such as longitudinal research designs. The increased popularity can be attributed to at least three factors. Linear mixed-effects models (LMEM) offer modeling flexibility in that the fixed effects, random effects, and the covariance structure can all be modeled. Also, parameters of LMEMs are estimated via maximum likelihood and hence have the asymptotic properties of being unbiased and efficient. In addition, because LMEM parameters are estimated through ML, the parameters can still be consistently estimated with missing data as long as the data are missing completely at random (MCAR) or missing at random (MAR)

Miguel A. Padilla is Assistant Professor of Quantitative Psychology. Email: mapadill@odu.edu. YoungKyoung Min is Senior Research Scientist. Email: ykymin@yahoo.com. Guili Zhang is Assistant Professor of Research and Evaluation Methodology. Email: zhangg@ecu.edu.

(Rubin, 1976). It is this last property which may ultimately account for the increased popularity of LMEMs. Even so, it is unclear exactly under which conditions LMEMs will have consistent parameter estimates when there are missing data.

When applying LMEM to split-plot designs, it is usually inferences about the fixed effects that are of main interest. Within this endeavor, a typical strategy is to try to fit a model for the means and select an appropriate covariance structure. The model is then tested for fit and appropriate modifications are made if required in order to test for inferences of interest (Wolfinger, 1993). A likelihood ratio, score, or Wald test can be used to test hypothesis about the fixed effect, but the Wald test is more commonly used (Schaalje, McBride, & Fellingham, 2002b; Brown & Prescott, 2006). The Wald test has good large sample properties, but they begin to dwindle with smaller sample sizes. However, using Satterthwaite-type degrees of freedom (Fai & Cornelius, 1996) can improve Wald test small sample properties. In addition to adjusting the degrees of freedom, the Wald test's small sample properties can further be enhanced by adjusting the covariance matrix (Kenward & Roger, 1997). Several simulation studies have shown that tests based on the Satterthwaite (SW) and Kenward-Roger (KR)

adjustments tend to behave well (Keselman et al., 1998; Schaalje, McBride, & Fellingham, 2002a; Padilla & Algina, 2007). In particular, the KR-test tends to behave well even with missing data (Padilla et al., 2007).

The small sample situation can further be complicated by missing data. It is a common occurrence in research and can have dramatic affects on the properties of standard statistical models, such as ordinary least squares regression. The way in which missing data will affect statistical models largely depends on the type of missing data mechanism and the way in which the missing data is handled. As an example, by far the most common method for handling missing data is to perform listwise deletion, also known as complete case analysis. This is most likely because it is the default in most popular statistical packages (e.g., SAS, SPSS, etc.). Nevertheless, if the data are MAR, parameter estimates can be biased and hence inference can be inaccurate. Additionally, there will be some loss of power in that participants with at least one missing value will be completely discarded from the analysis. If the small sample condition is added to this situation then the problems only worsen, adding another layer of uncertainty about inferences being drawn.

There are two major alternatives to handling missing data: multiple imputation (MI) and maximum likelihood (ML). Although both methods are a vast improvement over listwise deletion – and virtually any other method for handling missing data – the focus here will be on ML within the framework of split-plot designs and LMEMs. The reader interested in MI is referred to Schafer (1997) and Little & Rubin (2002).

The split-plot design is commonly used in behavioral research, such as educational and psychological research (Keselman et al., 1998b). It is, in essence, a hybrid of a between- and within-subjects designs incorporating elements of both. A longitudinal study is a typical split-plot design in that it has a between-subjects factor represented by subjects that are randomly assigned to treatment groups and a within-subjects factor represented by the measured multiple time points for each subject. Split-plot designs have various ways in which to analyze the data they generate and each of those methods have their strengths and limitations in terms of analyzing the data and how they handle missing values or data. However, the one promising technique for analyzing data from a split-plot with missing values is the linear mixed or mixed-effects model estimated through ML. Before delving on, the three missing data mechanisms are described.

Missing Data Mechanisms

The three general definitions of missing data, ordered from most restrictive to least restrictive, are missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR) (Rubin, 1976; Little & Rubin, 2002, p. 12). As described by Verbeke & Molenberghs (2000), let $f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right)$ denote the distribution of the missing data indicator or missing data mechanism for the i^{th} participant, where $\boldsymbol{r}_i$ is a $K \times 1$ vector containing zero for missing and one for observed scores in the corresponding $K \times 1$ $\boldsymbol{y}_i$ vector of repeated measurements or variables, $\boldsymbol{X}_i$ is the design matrix for the factors, and $\boldsymbol{\psi}$ contains the parameters of the relationship of $\boldsymbol{r}_i$ to $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$. Furthermore, $\boldsymbol{y}_i$ can be partitioned as $\boldsymbol{y}_i = \left(\boldsymbol{y}_{i(obs)}' \quad \boldsymbol{y}_{i(miss)}'\right)'$ where $\boldsymbol{y}_{i(obs)}'$ has observed scores and $\boldsymbol{y}_{i(miss)}'$ has missing scores for the i^{th} participant. The full data density can then be factorized as:

$$f\left(\boldsymbol{y}_i, \boldsymbol{r}_i \mid, \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\right) = f\left(\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}\right) f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right) \tag{1}$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \boldsymbol{\sigma}'\right)'$, $\boldsymbol{\beta}$ contains the fixed effects parameters, and $\boldsymbol{\sigma}$ contains the nonredundant parameters of the covariance matrix. This factorization is the foundation of selection modeling because the factor to the far right corresponds to the selection of individuals into observed or missing groups. The missing data are MCAR if $f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right) = f\left(\boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{\psi}\right)$, that is, the distribution of the missing data indicators does not depend on the repeated measures or variables. The missing data are

MAR if $f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right) = f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_{i(obs)}, \boldsymbol{X}_i, \boldsymbol{\psi}\right)$, that is, the distribution of the missing data indicator does not depend on the variables in which the i^{th} participant has missing scores. In general, missing data are NMAR if they are not MCAR or MAR. However, it is generally defined as

$f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{\psi}\right) = f\left(\boldsymbol{r}_i \mid \boldsymbol{y}_{i(miss)}, \boldsymbol{X}_i, \boldsymbol{\psi}\right)$, that is, the distribution of the missing data indicator depends on the missing values in the data.

A general method for consistent ML estimation of $\boldsymbol{\theta}$ is obtained by including both the missing data indicators ($\boldsymbol{r}_i$) and the parameters of their relationship to $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$ ($\boldsymbol{\psi}$) in the likelihood. The likelihood of the full data density can then be written as:

$$L\left(\boldsymbol{\theta}, \boldsymbol{\psi} \mid \boldsymbol{X}_i, \boldsymbol{y}_i, \boldsymbol{r}_i\right) \propto f\left(\boldsymbol{y}_i, \boldsymbol{r}_i \mid \boldsymbol{X}_i, \boldsymbol{\theta}, \boldsymbol{\psi}\right) \qquad (2)$$

If the missing data mechanism is MCAR or MAR and if $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are disjoint, ML estimators of $\boldsymbol{\theta}$ will be consistent if $\boldsymbol{r}_i$ and $\boldsymbol{\psi}$ are excluded from the analysis (Rubin, 1976). Dropping $\boldsymbol{r}_i$ and $\boldsymbol{\psi}$ is referred to as ignoring the missing data mechanism. Hence, MCAR or MAR missing data mechanisms are ignorable when model parameters ($\boldsymbol{\theta}$) are estimated via ML. If data are MCAR, listwise deletion and ML ignoring the missing data mechanism will produce consistent estimators, but ML estimators will be more precise because they use all available data.

In addition, Rubin (1976) showed that MCAR missing data mechanisms are ignorable for inferences based on sampling distributions. Thus, listwise deletion or ML ignoring the missing data mechanism can be used for inferences if the data are MCAR, but ML will result in more powerful inferences and narrower confidence intervals because it does not delete individuals with only partially observed scores on $\boldsymbol{y}_i$.

On the other hand, the validity of ML based inferences for a MAR missing data mechanism will depend on how the sampling covariance matrix is estimated. When the missing data mechanism is MAR, it will be ignorable if inferences are based on the sampling covariance obtained from the observed information matrix (Kenward & Molenberghs, 1998). This is in line with arguments from Efron & Hinkley (1978) in that the observed information matrix provides much better precision than the expected information matrix; that is, better variance component estimates. If ML inferences are based on the sampling covariance obtained from the expected information matrix, the MAR missing data mechanism may not be ignorable. The expected information matrix must take into account the actual sampling process implied by the MAR mechanisms in order for inferences to be valid (Kenward et al., 1998).

When the missing data mechanism is NMAR, then it is non-ignorable for purposes of ML estimation. In order to obtain consistent ML estimates in this particular case, the pattern of the missing values must be taken into account. A selection model that incorporates the missing values indicators ($\boldsymbol{r}_i$) or using a pattern mixture model that stratifies the data on the basis of the pattern of missing values can be used to obtain consistent ML estimates under an NMAR framework (Albert & Follmann, 2000; Diggle & Kenward, 1994; Fitzmaurice, Laird, & Shneyer, 2001; Kenward et al., 1998; Kenward, 1998; Troxel, Harrington, & Lipsitz, 1998; Algina & Keselman, 2004a; Algina & Keselman, 2004b; Little, 1995).

Linear Mixed-Effects Model

The linear mixed-effects model (LMEM) can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon} \qquad (3)$$

where $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are the design matrix and its corresponding fixed effects vector, $\boldsymbol{Z}$ and $\boldsymbol{u}$ are the design matrix and its corresponding random effects vector, and $\boldsymbol{\varepsilon}$ is the vector of random errors. It is generally assumed that $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are independent, hence

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}\right) \qquad (4)$$

Based on this assumption, $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ and $Var(\boldsymbol{y}) = \boldsymbol{V} = \boldsymbol{ZGZ'} + \boldsymbol{R}$. A common estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{y} \quad (5)$$

Also, $Var\left(\hat{\boldsymbol{\beta}}\right) = \left(\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ is the estimated generalized least-squares covariance of $\hat{\boldsymbol{\beta}}$.

Let $\boldsymbol{L}$ be a contrast matrix of full row rank *r*. Then the main effect and interaction hypothesis about the between- and within-subjects factors can be expressed as $H_0 : \boldsymbol{L\beta} = \boldsymbol{0}$. The common test statistic for this hypothesis is the Wald

$$F_{r,ddf} = \frac{\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{L}\left(\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{L'}\right)^{-1}\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)}{r} \quad (6)$$

where *ddf* is the denominator degrees of freedom. It should be noted that, under the null hypothesis, the Wald $F_{r,ddf}$ approximately follows an *F* distribution. However, there are times when it follows an *F* distribution exactly. Even so, when there is no missing data, $\left(\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ tends to underestimate $\left(\boldsymbol{X'}\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}$ and hence is a biased estimate because it fails to take into account the uncertainty introduced by using $\hat{\boldsymbol{V}}$ (Booth & Hobert, 1998; Kackar & Harville, 1984; Prasad & Rao, 1990).

Kenward-Roger *F*-Test

Better estimates were developed as a response to the poor statistical properties of $Var\left(\hat{\boldsymbol{\beta}}\right)$. The first estimate, denoted as $Var\left(\hat{\boldsymbol{\beta}}^{@}\right) = \hat{\boldsymbol{m}}^{@}$, was proposed by Harville & Jeske (1992). Subsequently, Kenward & Roger (1997) developed an alternative estimator, denoted as $Var\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{A}}\right) = \hat{\boldsymbol{\Phi}}_A$. Additionally, Kenward & Roger derived the test statistic

$$F^*_{r,d} \simeq \lambda\frac{\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{L}\hat{\boldsymbol{\Phi}}_A\boldsymbol{L'}\right)^{-1}\left(\boldsymbol{L}\hat{\boldsymbol{\beta}}\right)}{r} \quad (7)$$

where λ is a scaling factor and *d* is the approximate denominator degrees of freedom. As in the case of $F_{r,ddf}$, $F^*_{r,d}$ is assumed to follow an *F* distribution under the null hypothesis. Both λ and *d* are calculated from the data. First, $\hat{\boldsymbol{\Phi}}_A$ is estimated to account for small sample bias in $\left(\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ and variability introduced by using $\hat{\boldsymbol{V}}$ (Kackar et al., 1984). Then *d* is approximated by using the spectral decomposition of $\left(\boldsymbol{L}\hat{\boldsymbol{\Phi}}_A\boldsymbol{L'}\right)^{-1}$ concurrently with repeated applications of the single degree of freedom *t*-test (Fai et al., 1996; Giesbrecht & Burns, 1985). The Kenward-Roger (KR) *F*-test is implemented in SAS PROC MIXED, but uses $\hat{\boldsymbol{m}}^{@}$ instead of $\hat{\boldsymbol{\Phi}}_A$. (See Padilla & Algina, 2007) for how to specify model parameters using the mean vector and an indicator matrix for the missing values.)

Some research has been conducted investigating the Type I error rate of the KR method (Fai et al., 1996; Kenward & Roger, 1997; Kowalchuk, Keselman, Algina, & Wolfinger, 2004; Gomez, Schaalje, & Fellingham, 2005). However, very little research is available on the Type I error rate of the KR method when there are missing values. To date, Padilla & Algina (2007) is the only work investigating the Type I error rate of the KR *F*-test when the missing values are MAR.

Fai & Cornelius (1996) derived four test statistics (F_1, F_2, F_3, F_4) for hypothesis testing on the means in multivariate data. The F_1 and F_2 statistics use $\left(\boldsymbol{X'}\hat{\boldsymbol{V}}^{-1}\boldsymbol{X}\right)^{-1}$ whereas F_3 and F_4 use $\hat{\boldsymbol{m}}^{@}$ to estimate $Var\left(\hat{\boldsymbol{\beta}}\right)$. Additionally, F_2 and F_4 have scaling factors λ_2 and λ_4, respectively. The F_1 statistic is available in SAS PROC MIXED when the Satterthwaite option is used for DDFM. The F_4 statistic is similar to the PROC MIXED KR *F*-test, but uses a different formula for the scaling factor and denominator

degrees of freedom. (See Fai & Cornelius for further details.)

Fai & Cornelius (1996) applied their tests to simulated data from four unbalanced 3 (between) × 4 (within) split-plot designs with a compound symmetric covariance structure. Imbalance was created by varying the number of subjects of the between-subjects factor without generating some combinations of subjects and the within-subjects factor. Missing data were never actually generated; hence the missing data mechanism is MCAR. The four unbalanced designs had total sample sizes of $N = 25, 34, 40, 48$. Estimated Type I error rate and power were reported for the between-subjects main effects. All tests controlled the Type I error rate reasonably well. The results of F_1 and F_3 were similar, and power and Type I error were always larger for F_4 than for F_3.

In their initial work, Kenward & Roger (1997d) investigated the Type I error rate of the KR *F*-test in simulated data from four research designs: (a) a four-treatment, two-period cross-over, (b) a row-column-α design, (c) a random coefficients regression model for repeated measures data, and (d) a split-plot design. Design (c) had MCAR missing values and (d) had missing values with an unspecified missing data mechanism. Estimated Type I error rates were reported for the between-subjects main effect. In all situations, the KR *F*-test Type I error rate was well controlled.

Kowalchuk, Keselman, Algina, & Wolfinger (2004) compared the Type I error rates of the KR and Welch-James (WJ) *F*-tests under several simulation conditions for a 3 (between) × 4 (within) split-plot design. Investigated conditions were (a) type of covariance structure, (b) group size inequality, (c) positive and negative parings of covariance matrices with group sample sizes, (d) shape of data distribution, and (e) type of covariance structure fit to data. A heterogeneous covariance structure with a 1:3:5 ratio was used for all simulation conditions, and missing values were not investigated. Estimated Type I error rates were reported for the main effects and interaction. Under all conditions with small sample sizes (total N = 30, 40), the Type I error rate of the KR *F*-test were closer to the target value ($\alpha = .05$) than the WJ *F*-test. Additionally, the Type I error rates of the KR *F*-test were always comparable when using an unstructured covariance matrix to modeling the true covariance matrix.

Gomez, Schaalje, & Fillingham (2005) investigated the Type I error rate of the KR *F*-test when using AIC (Akaike, 1974) and BIC (Schwarz, 1978) to select the covariance structure. Investigated conditions were (a) type of covariance structures with within- and between-subjects heterogeneity (1:3:5 ratio for between-subjects), (b) equal ($total\ N = 9, 15$) and unequal group sample sizes ($n = 3, 5, 7$), (c) positive and negative paring for unequal group sample sizes, (d) and levels of the within-subjects factor ($K = 3, 5$). The between-subjects factor was fixed at 3 and no missing values were investigated. Estimated Type I error rates were reported for the main effects only. In general the Type I error rate was close to the target value when the correct covariance structure was used. However, the Type I error rate becomes inflated with complex covariance structures and small sample sizes. Additionally, the Type I error rate increased with heterogeneity within- and between-subjects, and even more so with negative pairings. In general, the success rate of choosing the correct covariance structure was low for both the AIC and BIC. At most, the success rate was 73.91%. Even so, the success rate was higher for the larger sample sizes and simpler covariance structures. Lastly, the AIC had better success with complicated covariance structures and the BIC with simpler ones.

Padilla & Algina (2007) studied the Type I error rate of the KR *F*-test with missing values and heterogeneity of covariance matrices ($1:3:5\ ratio$). Investigated conditions were (a) level of between-subject factor (*J*), (b) level of within-subject factor (*K*), (c) $n_{\min}/(K-1)$, (d) sample size inequality, (e) degree of sphericity, (f) covariance and group sample size pairing, (g) missing data mechanism (MCAR or MAR), and (h) percent of missing data. Estimated Type I error rates were reported for the main effects and interaction. In general, the Type I error rates of

the KR *F*-test were close to the target value of $\alpha = .05$ for the between- and within-subjects main effects and the between- by within-subjects interaction. The best Type I error control was attained by the between-subjects main effect with the between- by within-subjects interaction attaining the worst. However, the distribution of the data was normal.

The previous studies demonstrate that the Type I error rate of the KR *F*-test remains close to the target value $(\alpha = .05)$ under a variety of repeated measures designs and simulation conditions, which included MCAR unbalanced data. However, Padilla & Algina (2007) is the only study to investigate the Type I error rate of the KR *F*-test under the MAR condition in normal data. This study builds on Padilla & Algina and investigates the Type I error rate of the KR *F*-test under several simulation conditions. Of particular interest is the KR *F*-test Type I error rate when data are non-normal with missing values as it is implemented in SAS PROC MIXED.

Methodology

Design

The simplest of the split-plot design with one between- and one within-subjects factor (*i.e.*, $J \times K$) with heterogeneity between the j^{th} covariance matrix and non-normal data was investigated. In this type of design subjects are randomly assigned to the levels of the between-subjects factor $\left(j = 1, 2, \ldots, n; \sum_j n_j\right)$ and measured under all levels of the within-subjects factor $(k = 1, 2, \ldots, K)$. The heterogeneity between the j^{th} covariance matrices was set at 1:3:5; that is $\boldsymbol{\Sigma}_1 = 1/3\,\boldsymbol{\Sigma}_2$ and $\boldsymbol{\Sigma}_3 = 5/3\,\boldsymbol{\Sigma}_2$ (Algina & Keselman, 1997; Keselman, Algina, Kowalchuk, & Wolfinger, 1999; Padilla et al., 2007; Keselman, Carriere, & Lix, 1993). The non-normal data were generated from a multivariate lognormal distribution under the null using the methods outlined in Algina & Oshima (1994) with skewness set at 1.75 and kurtosis at 5.90 (Keselman, Algina, Wilcox, & Kowalchuk, 2000; Kowalchuk, Keselman, Algina, & Wolfinger, 2004).

All simulations and analyses were done on SAS 9.1. The PROC MIXED code for estimating the Kenward-Roger *F*-test can be found in Padilla and Algina (2007).

Simulation Variables

Eight variables were investigated. The variables of interest are (a) number of levels of the between-subjects factor (*J*), (b) number of levels of the within-subjects factor (*K*), (c) sample size, (d) sample size inequality across the j^{th} groups, (e) degree of sphericity, (f) pairing of the j^{th} group sizes with covariance matrices, (g) type of missing data, and (h) percent of missing data. Because this study builds on Padilla & Algina (2007), the simulation variables here are similar to theirs.

Between- and Within-Subjects Factors

The between- and within-subjects factors each had two levels with $J, K = 3, 6$.

Sample Size

Sample sizes were based on the $n_{\min}/(K-1)$ ratio (Keselman, Carriere, & Lix, 1993b). The ratios were set as in Padilla & Algina (2007) and for the same reasons. The actual sample sizes used, in combination with sample size inequality, are displayed in Tables 1 and 2.

Table 1:
Groups Sizes for Each Level of *J* at *K* = 3

	Sample Size Inequality			
J	C ≈ .16	C ≈ .33	C ≈ .16	C ≈ .33
	$n_{\min}/(K-1) = 4.0$		$n_{\min}/(K-1) = 6.0$	
3	8	8	12	12
	10	14	15	20
	12	20	18	28
	$n_{\min}/(K-1) = 5.0$		$n_{\min}/(K-1) = 7.7$	
6	10	10	15	15
	13	17	19	25
	16	24	23	35
	10	10	15	15
	13	17	19	25
	16	24	23	35

Table 2:
Groups Sizes for Each Level of J at $K = 6$

	Sample Size Inequality			
J	C ≈ .16	C ≈ .33	C ≈ .16	C ≈ .33
	$n_{min}/(K-1) = 4.0$		$n_{min}/(K-1) = 6.0$	
	20	20	30	30
3	25	34	37	50
	30	48	44	70
	$n_{min}/(K-1) = 5.0$		$n_{min}/(K-1) = 7.7$	
	25	25	38	38
	31	42	47	64
6	37	59	56	90
	25	25	38	38
	31	42	47	64
	37	59	56	90

Sample Size Inequality

Unequal sample sizes are common in split-plot designs and hence were investigated here (Keselman et al., 1998). The unequal group sample size were investigated through the coefficient of variation as defined by Keselman et al. (1993):

$$C = \left(\bar{n}\sqrt{J}\right)^{-1}\sqrt{\sum_{j=1}^{J}\left(n_j - \bar{n}\right)^2} \qquad (8)$$

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

Covariance Sphericity

Sphericity as quantified by Box's epsilon (1954) was investigated with $\varepsilon = .60, .75, .90$. Here, $\varepsilon = .60$ represents a relatively severe departure from sphericity whereas $\varepsilon = .75$ a moderate one. Epsilon values were chosen based on the argument that $\varepsilon = .75$ represent the lower limit of ε found in educational and psychological data (Huynh & Feldt, 1976). (See Padilla & Algina (2007) for the actual covariance matrices.)

Group Pairing with Covariance

Pairing of the unequal group samples sizes and heterogeneous covariance matrices were investigated. The two conditions investigated were positive and negative pairings because positive pairing tend to produce conservative Type I error rates whereas negative pairings tend to produce liberal ones (Keselman & Keselman, 1990). A positive pairing occurs when the largest n_j is paired with the covariance matrix with the largest elements and a negative pairing occurs when the largest n_j is paired with the covariance matrix with the smallest elements. For positive pairings, the ratios of group sample size to heterogeneity of covariance matrices was set at $5:3:1$ for $J = 3$ and $5:3:1:5:3:1$ for $J = 6$. For negative pairings, it was set at $1:3:5$ for $J = 3$ and $1:3:5:1:3:5$ for $J = 6$.

Missing Data Mechanism

Both MCAR and MAR missing data mechanisms were investigated. The missing data mechanisms were simulated as described by Padilla & Algina (2007). NMAR was not investigated because it negatively impacts the Type I error rate of the KR F-test in a repeated measures designs with no between-subjects factor and normal data (Padilla & Algina, 2004).

Percent of Missing Data

Five percent (5%) and 15% probability of missing data at each level of the within-subjects factor were investigated. The exception here is that there was no missing data in the first level. Higher missing data probabilities were not investigated because the sample sizes are considerably small (see Table 1) and this will impede the convergence of the Newton-Raphson algorithm.

Analysis

The p-values of KR F-test were available from 5,000 replications for each combination of the simulation variables. The Type I error for each of the p-values was defined as

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

Logistic regression models were used to analyze the between-subjects main effect, within-subjects main effect, and the between- by within-subjects interaction of the KR *F*-test separately. In each logistic model the Type I error variable was used as the dependent variable with the simulation variables as the independent variables. A forward selection approach was used to select appropriate models beginning with the intercept-only model and moving up to main effect only, main effect with two-way interaction, etc. A model adequately fit the data if the χ^2 goodness of fit test was non-significant or if $CFI \geq .95$ (Bentler, 1990). With large sample sizes (i.e., number of replications), the χ^2 goodness of fit statistic is sensitive to small effects, hence a fit index was used to supplement the χ^2. In this context, the CFI is calculated as follows:

$$CFI = 1 - \left(\lambda / \lambda_i\right) \tag{9}$$

where $\lambda = \max\left(\chi^2 - df, 0\right)$ with χ^2 (the test statistic) and *df* (the degrees of freedom) for the fitted model and $\lambda_i = \max\left(\chi_i^2 - df_i, \chi^2 - df, 0\right)$ with χ_i^2 and df_i for the intercept-only model.

Bradley's (1978) liberal criterion was used to assess the Type I error rates. The liberal criterion is $.5\alpha \leq \tau \leq 1.5\alpha$ where α is the nominal Type I error and τ is the empirical Type I error. With $\alpha = .05$ the liberal range is $.025 \leq \tau \leq .075$. Hence if the Type I error is within the range, the test is considered to be robust.

Results

Between-Subjects Main Effect

The logistic model with main effects and two-way interactions had $\chi^2(339) = 388.40$, $p = .0331$ and $CFI = .98$. Inspection of all two-way interaction tables indicated that for the between-subjects main effects all Type I error rates were within Bradley's liberal criterion. In fact the range of the Type I error rates across all two-way interaction tables was [.051, .071]. Even though the KR *F*-test for the between-subjects main effect does appear to be slightly liberal, it is not too strongly affected by the simulation variables.

Within-Subjects Main Effect

The logistic model with main effects and three-way interactions had $\chi^2(262) = 261.76$, $p = .4925$ and $CFI = 1.00$. Therefore, the three-way interaction model was selected for further analysis. Wald tests of the logistic model indicated that levels of the *within-subjects factor* (*K*), *group pairing with covariance*, *missing data mechanism*, and *percent of missing data* had significant main effects and also entered into the most significant three-way interactions.

Mean Type I error rates are displayed in Table 3. The range of mean Type I error rates under MCAR was [.054, .067]. Although slightly liberal, the mean Type I error rates are well within Bradley's liberal criterion. Under MAR, the situation changes dramatically. In fact, the mean Type I error rates were all liberal ranging from [.079, .158] and above Bradley's liberal criterion. Furthermore, the mean Type I error rate increases as both the levels of the *within-subjects factor* (*K*) and *percent of missing data* increases. On the other hand, under MAR, the mean Type I error rate decreases as the *group pairing with covariance* changes from positive to negative (consistent with Keselman et al., 1990).

Table 3: Within-Subiects Main Effect

Missing Data Mechanism	% Missing	*K*	Group Pairing: Positive	Group Pairing: Negative
MCAR	5	3	.0625	.0670
		6	.0543	.0572
	15	3	.0634	.0670
		6	.0607	.0631
MAR	5	3	**.0794**	**.0794**
		6	**.0938**	**.0880**
	15	3	**.1078**	**.0986**
		6	**.1580**	**.1389**

Note: Type I error rate above Bradley's liberal criterion are in bold type.

Between- by Within-Subjects Interaction

The logistic model with main effects and three-way interactions had $\chi^2(262) = 308.64$, $p = .0252$ and $CFI = 1.00$. Hence, the three-way interaction model was selected for additional analysis. Wald tests of the logistic model indicated that *K*, *J*, *sample size*, *group pairing with covariance*, *covariance sphericity*, and *percent of missing data* had significant main effects. However, *K*, *J*, *sample size inequality*, *group pairing with covariance*, *missing data mechanism*, and *percent of missing data* entered into the most significant three-way interactions. Thus, these latter simulation variables were selected for further analysis.

Mean Type I error rates under MCAR are displayed in Table 4. With the exception of 15% missing data, a negative pairing, and a severe group sample size inequality, the majority of mean Type I error rates are within Bradley's liberal criterion. However, the mean Type I error rates increase as the *percent of missing data*, *K*, and *J* increases and as *group pairing* changes from positive to negative. As noted above, the situation becomes more aggravated under the most severe conditions of the simulation variables.

Mean Type I error rates under MAR are presented in Table 5. Here, most of the mean Type I error rates are outside of the range of the Bradley's liberal criterion. The only time the mean Type I error rate is controlled is under the simplest of conditions for *group pairing with covariance*, *K*, and *J*. Nevertheless, as was the case for the MCAR condition, the mean Type I changes from positive to negative. The one difference is that mean Type I error rate error rates tend to increase as *percent of missing data*, *K*, and *J* increases and as *group pairing* increases as the *sample size inequality* becomes more severe. Not surprising the mean Type I error rates become more liberal under the more severe conditions of the simulation variables.

Conclusion

The results indicate that sampling distribution based inferences on the means for the between-subjects factor of a split-plot design using ML estimates can control the Type I error rate under an MCAR and MAR missing data mechanism and non-normal data. Furthermore, the Type I error control can be achieved with relatively small to moderate sample sizes when using the KR *F*-test. The same cannot be said of inferences about the within-subjects factor or the within- by between-subjects interaction.

The Type I error rates of the KR *F*-test for the latter two cases are impacted by several conditions of the simulation variable with the most dramatic being the MAR condition. This is most clearly seen in inferences about the within-subjects factor, in which case none of the Type I error rates were acceptable. Under MCAR, increasing the percent of missing data and switching from a positive to negative pairing of groups with covariance matrices tended to increase the Type I error rate, but the Type I error rate was still within Bradley's (1978) liberal criterion. Although the same pattern of increase in Type I error rate is observed under MAR, the increase in Type I error rate was

Table 4: MCAR for Interaction

% Missing	Group Pairing	Sample Size Inequality	*K* = 3		*K* = 6	
			J = 3	*J* = 6	*J* = 3	*J* = 6
5	Positive	Moderate	.0495	.0575	.0446	.0549
		Severe	.0468	.0527	.0456	.0513
	Negative	Moderate	.0642	**.0769**	.0569	.0607
		Severe	**.0787**	**.0864**	.0614	.0673
15	Positive	Moderate	.0503	.0582	.0562	.0584
		Severe	.0517	.0564	.0542	.0629
	Negative	Moderate	.0679	**.0774**	.0637	.0709
		Severe	**.0781**	**.0910**	.0715	**.0769**

Note: Type I error rate above Bradley's liberal criterion are in bold type.

Table 5: MAR for Interaction

% Missing	Group Pairing	Sample Size Inequality	*K* = 3		*K* = 6	
			J = 3	*J* = 6	*J* = 3	*J* = 6
5	Positive	Moderate	.0504	.0637	.0665	**.0859**
		Severe	.0523	.0661	.0728	**.0994**
	Negative	Moderate	.0672	**.0843**	.0696	**.0977**
		Severe	**.0823**	**.1041**	**.0820**	**.1181**
15	Positive	Moderate	.0561	**.0789**	**.0987**	**.1582**
		Severe	.0668	**.0963**	**.1247**	**.2009**
	Negative	Moderate	.0699	**.0991**	**.0954**	**.1665**
		Severe	**.0916**	**.1338**	**.1200**	**.2208**

Note: Type I error rate above Bradley's liberal criterion are in bold type.

sharper and obvious when switching from MCAR to MAR in which case none of the Type I error rates were within Bradley's liberal criterion.

With regard to the within- by between-subjects interaction, the KR *F*-test is once again severely impacted by several of the simulation conditions, but more dramatically by the MAR condition. Under the MCAR condition the majority of the Type I error rates are within Bradley's liberal criterion. When the within-subjects factor is 3, the same pattern is observed for 5% and 15% missing data: a negative pairing of groups with covariance matrices coupled with severe sample size inequality increased the Type I error rate above the liberal criterion. When the within-subject factor is 6, the Type I error rate was above the liberal criterion only under the more severe simulation conditions. Under MAR, most of the Type I error rates were above the liberal criterion. The only time the Type I error rates were consistently within the liberal criterion was when the within-subjects factor was 6, 5% of the data were missing, and there was positive pairing of groups with covariance matrices. The remaining acceptable Type I error rates tended to occur when the between-subjects factor was 3 and under the least severe of the simulation conditions. Even so, the Type I error rate tended to increase as the simulation conditions switched into the more severe conditions investigated.

By far the MAR condition had the largest impact on the Type I error rate of the KR *F*-test for the within-subjects factor and the within- by between-subjects interaction. It is clear that missing values coupled with non-normal data impact the accuracy of the *F*-distribution as an approximation to the sampling distribution of the KR *F*-test. The KR *F*-test uses an adjusted estimator of the covariance which is then used to estimate Satterthwaite type degrees of freedom. The procedure provides a better approximation to the *F*-distribution with small sample sizes (Kenward & Roger, 1997). This seemed to be the case for the between-subjects factor under all the simulation variables of this study. However, for the within-subjects factor and the within- by between-subjects interaction, it appears that the MAR condition coupled with non-normal data severely limited the KR *F*-test's ability to control the Type I error.

Two potential reasons exist for this result. First, SAS PROC MIXED does not compute the covariance matrix by inverting the Hessian (information matrix) for the fixed effects and the covariance parameters. According to Verbeke & Molenberghs (2000), the observed Hessian should be used and not the expected Hessian. Again, the observed Hessian provides more precision than the expected Hessian (Efron & Hinkley, 1978). Second, sample sizes were too small; particularly when

the within- and between-subjects factors were both set at six. Although the samples sizes were based on the recommendations set by Keselman et al. (1993a) and Algina & Keselman (1997), those studies did not have missing values, which is not the case here. Here it appears that missing values coupled with data non-normality put a heavy burden on the analysis. A simple solution is to increase the sample sizes. However, doing so will increase the computation time of PROC MIXED's KR procedure, but it should provide more information for the procedure to use. However, increasing the sample sizes is not easy in practice.

The KR *F*-test for the between-subjects factor appears to be robust, in terms of controlling the Type I error, to non-normal data under the simulation variables investigated. Also, the KR *F*-test for the within-subjects factor is robust to non-normal data under the simulation variables investigated as long as the missing data mechanism is MCAR. The KR *F*-test for the within- by between-subjects interaction performed fairly well under MCAR, but care should be taken when using it when the within-subjects factor is three and the more extreme conditions of the simulation variables. Unfortunately, the KR *F*-test for the within- by between-subjects interaction is not robust under MAR and the simulation variables investigated. The only time the KR *F*-test for the interaction provided acceptable Type I error rates was when the within-subjects factor was set at 3 and only 5% of the data were missing. More work is required in order to fully assess the KR *F*-test's Type I error rate under missing values and non-normal data.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, *19*, 716-723.

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

Algina, J., & Keselman, H. J. (1997). Testing repeated measures 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. (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 general approximation and improved general approximation tests. *British Journal of Mathematical & Statistical Psychology*, *47*, 151-165.

Booth, J. G., & Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, *93*, 262-272.

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

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

Brown, H., & Prescott, R. (2006). *Applied mixed models in medicine*. (2nd ed.) Wiley: New York.

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

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*, 457-481.

Fai, A. H. T., & 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 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.

Giesbrecht, F. G., & 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.

Gomez, E. V., Schaalje, G. B., & Fellingham, G. W. (2005). Performance of the Kenward-Roger method when the covariance structure is selected using AIC and BIC. *Communications in Statistics-Simulation and Computation*, *34*, 377-392.

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. (1976). Estimation of the Box correction for degrees 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, B. K., & Wolfinger, R. D. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical & Statistical Psychology*, *52*, 63-78.

Keselman, H. J., Algina, J., Wilcox, R. R., & Kowalchuk, R. K. (2000). Testing repeated measures hypotheses 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-measures 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., Donahue, B., et al. (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). Analyzing unbalanced repeated measures designs. *British Journal of Mathematical and Statistical Psychology*, *43*, 265-282.

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.

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*. (2nd ed.) New York: Wiley.

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.

Padilla, M. A., & Algina, J. (2007). Type I error rates of the Kenward-Roger adjusted degree of freedom F-test for a split-plot design with missing values. *Journal of Modern Applied Statistical Methods*, *6*.

Prasad, N. G. N., & Rao, J. N. K. (1990). The estimation of the 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.

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

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

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464.

Troxel, A. B., Harrington, D. P., & Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Journal of the Royal Statistical Society Series C-Applied Statistics*, *47*, 425-438.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer-Verlag.

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics-Simulation and Computation*, *22*, 1079-1106.