Comparison of Denominator Degrees of Freedom Approximations

for Linear Mixed Models in Small-Sample Simulations

by

Ping-Chieh Huang

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science in Statistics

Approved April 2020 by the
Graduate Supervisory Committee:

Mark Reiser, Chair
Ming-Hung Kao
Jeffrey Wilson

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

Whilst linear mixed models offer a flexible approach to handle data with multiple sources of random variability, the related hypothesis testing for the fixed effects often encounters obstacles when the sample size is small and the underlying distribution for the test statistic is unknown. Consequently, five methods of denominator degrees of freedom approximations (residual, containment, between-within, Satterthwaite, Kenward-Roger) are developed to overcome this problem. This study aims to evaluate the performance of these five methods with a mixed model consisting of random intercept and random slope. Specifically, simulations are conducted to provide insights on the F-statistics, denominator degrees of freedom and p-values each method gives with respect to different settings of the sample structure, the fixed-effect slopes and the missing-data proportion. The simulation results show that the residual method performs the worst in terms of F-statistics and p-values. Also, Satterthwaite and Kenward-Roger methods tend to be more sensitive to the change of designs. The Kenward-Roger method performs the best in terms of F-statistics when the null hypothesis is true.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

1 Background of Study

1.1 Introduction to Linear Mixed Models

Linear mixed models (LMMs) are useful in analyzing data with multiple sources of random variability and particularly handy in settings where repeated measurements are made in a longitudinal study. LMMs are an extension of standard linear models, involving a mixture of linear functions of fixed effects and random effects. These two types of effects are distinguished based on how they change among observations. While fixed effects remain constant, random effects may vary across observations (Kreft and De Leeuw, 1998). Such random variation is often addressed by the serial correlation and the cluster correlation. Serial correlation is present when the units are repeatedly measured over a time- or space-varying stochastic process (Diggle et al., 2002). Cluster correlation occurs when the observations are grouped in a variety of ways, such as repeated random sampling of subgroups or repeated measuring of the same units (Rencher and Schaalje, 2008). The general setting of a linear mixed model can be expressed in a matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \varepsilon,$$

where $\mathbf{y}$ is a vector of responses with mean $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{X}$ is a matrix of known constants for fixed effects, $\boldsymbol{\beta}$ is a vector of unknown fixed effects, $\mathbf{Z}$ is a matrix of known constants for random effects, $\mathbf{u}$ is a vector of unknown random effects with $\mathbf{u} \sim \mathrm{N}[\mathbf{0},\ \mathbf{G}(\boldsymbol{\theta}_1)]$, ε is a vector of unknown random errors with $\varepsilon \sim \mathrm{N}[\mathbf{0},\ \mathbf{R}(\boldsymbol{\theta}_2)]$, and $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are vectors of variance parameters. Assuming $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1{}',\ \boldsymbol{\theta}_2{}')$, the covariance matrix of $\mathbf{y}$ is

$\boldsymbol{\Sigma} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.

The parameters in the mixed model can be estimated by either the maximum likelihood (ML) approach or the restricted maximum likelihood (REML) approach. In the case of large samples, the ML approach is usually robust against mild violations of the assumptions and gives estimators that are asymptotically consistent and efficient (Hox et al., 2018). However, when estimating the variance components, the ML approach does not take into account the loss of degrees of freedom resulting from the estimation of the fixed effects (West et al., 2015). The result is that the ML estimators are biased with smaller variances, especially when the sample size is small (Searle et al., 2006). On the other hand, the REML approach, taking into account the loss of degrees of freedom, is considered a better way with respect to the estimation of the variance components (Snidjers and Bosker, 1999). The REML estimators of the variance components are also invariant to the value of $\boldsymbol{\beta}$ and less sensitive to the outliers in the data, compared to the ML estimators (McCulloch et al., 2008). In view of these preferable characteristics, the REML approach is chosen over the ML approach in this study and will be adopted into the simulations in the later chapter.

1.2 Issues with Inferences for Fixed Effects

The REML approach (Patterson and Thompson, 1971; Harville, 1977) provides essential estimators that are used to derive the F-test statistic for the fixed effects in the mixed model. $\hat{\boldsymbol{\Sigma}}$, an estimator of the covariance matrix $\boldsymbol{\Sigma}$, is the inverse of the Hessian matrix of the restricted log likelihood function. Given $\hat{\boldsymbol{\Sigma}}$, the estimated generalized least squares estimator of $\boldsymbol{\beta}$ can be obtained as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$ and if $\mathbf{X}$ is full rank the approximate covariance matrix of $\hat{\boldsymbol{\beta}}$ can be expressed as $\mathrm{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}$.

Extending the theorem of the general linear hypothesis test (Rencher and Schaalje, 2008), we obtain that for a known full-rank $q \times p$ matrix $\mathbf{L}$ whose rows define the estimable functions of β, $\mathbf{L}\hat{\beta}$ approximately distributes as $\mathrm{N}_q[\mathbf{L}\beta, \mathbf{L}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-}\mathbf{L}']$ and $(\mathbf{L}\hat{\beta} - \mathbf{L}\beta)'[\mathbf{L}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-}\mathbf{L}']^{-1}(\mathbf{L}\hat{\beta} - \mathbf{L}\beta)$ is approximately $\chi^2(q)$. Further with that (1) $[\mathbf{L}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-}\mathbf{L}']^{-1}$ is formed as $\frac{d}{w}\mathbf{K}$, where w is a central chi-square random variable with d degrees of freedom, and that (2) $(\mathbf{L}\hat{\beta} - \mathbf{L}\beta)'\mathbf{K}(\mathbf{L}\hat{\beta} - \mathbf{L}\beta)$ follows a (most-likely noncentral) chi-square distribution with q degrees of freedom; a test statistic for the hypothesis $\mathrm{H}_0 : \mathbf{L}\beta = \mathbf{t}$ can be derived as an F-test statistic (Rencher and Schaalje, 2008),

$$F = \frac{(\mathbf{L}\hat{\beta} - \mathbf{t})'\mathbf{K}(\mathbf{L}\hat{\beta} - \mathbf{t})/q}{w/d} = \frac{(\mathbf{L}\hat{\beta} - \mathbf{t})'[\mathbf{L}(\mathbf{X}'\hat{\Sigma}^{-1}\mathbf{X})^{-}\mathbf{L}']^{-1}(\mathbf{L}\hat{\beta} - \mathbf{t})}{q}.$$

Nevertheless, these inferences are not universally satisfactory for small samples. In most cases of small samples, especially with complex models consisting of unbalanced datasets and complicated covariance structures, the distribution of the test statistic is unknown and the p-values cannot be computed exactly. In this regard, the estimation of the denominator degrees of freedom (DDF) becomes the most critical issue on conducting an approximate F-test and providing informative inferences (Schaalje et al., 2002). As of today, a number of DDF approximation methods have been developed (Schluchter and Elashoff, 1990; Fai and Cornelius, 1996; Kenward and Roger, 1997) and a variety of research has been done in examining the performance of these methods (Tietjen, 1974; Li and Redden, 2015; Luke, 2016). Yet, besides the variability inherent in the data and the research questions, the development of software packages assisting the mixed-model computation further brings uncertainty and unknown into research and practice. Therefore more studies using software packages on how the DDF approximation methods perform under different research designs especially with complex settings are required to provide insights for future theoretical development and industrial practice.

In this study, we aim to investigate the five commonly-used DDF approximation methods (residual, containment, between-within, Satterthwaite, and Kenward-Roger) that are available in SAS and to conduct simulations to evaluate their performance with a relatively complicated model and different sample structure and missing-data proportion. Specifically, in Chapter 2 we review the theoretical background of each method to lay the foundations for our study; in Chapter 3 we present the results of our simulations with a mixed model including random intercept and random slope and under the effects of imbalance and sample structure; and finally in Chapter 4 we conclude with a discussion of our results and the area worth further exploration.

2 Denominator Degrees of Freedom Approximations

2.1 Residual Method

The residual method is the simplest method, in which all of the tests are performed using the residual degrees of freedom $\mathrm{N} - \mathrm{rank}(\mathbf{XZ})$, where N is the total sample size, **X** is the matrix of known constants for fixed effects, and **Z** is the matrix of known constants for random effects. Only for independent and identically distributed designs this method provides correct denominator degrees of freedom; ignoring the covariance structure of the model, this method only performs well in large-sample situations where the asymptotic distributions provide good approximations (Schaalje et al., 2002). Although in this study our main focus is correlated data in small-sample situations, we include this method in our simulations to see how the F-statistics, DDF, and p-values it gives may differ from the other methods.

2.2 Containment Method

The containment method is to search through the random effects that syntactically contain the fixed effects of interest (based on the statement syntax defining the random effects), compute their contributions to the $\mathrm{rank}(\mathbf{XZ})$, and assign the minimum of these rank contributions as the DDF. If no such random effects can be found or no random effects are clearly stated to contain the fixed effects of interest, the DDF is set equal to the residual degrees of freedom (West et al., 2015). Schaalje et al. (2002) note that the containment method is considered to give exact DDF when the design is balanced, there is no structure in the **R** matrix (the covariance matrix of unknown random errors), and the nested fixed-random syntax is clearly stated; otherwise this method may provide adequate approximate results.

2.3 Between-Within Method

The between-within (B-W) method proposed by Schluchter and Elashoff (1990) is to divide the residual degrees of freedom into between-subject and within-subject portions, and then assign the DDF: (1) by the within-subject degrees of freedom, if the fixed effects change within any subject; or (2) by the between-subject degrees of freedom, otherwise. Exceptionally for cases that multiple within-subject effects including classification variables are present, the within-subject degrees of freedom are further partitioned into components that correspond to the subject-by-effect interactions (SAS Institute, 2017). The between-within method is considered to provide exact DDF for balanced repeated-measure designs involving a covariance structure of compound symmetry; in other cases, this method may give approximate results at best and can be unpredictable sometimes (Schaalje et al., 2002).

2.4 Satterthwaite Method

The Satterthwaite method is to obtain an approximate small-sample distribution of $F = \frac{(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{t})'[\mathbf{L}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{L}']^{-1}(\mathbf{L}\hat{\boldsymbol{\beta}} - \mathbf{t})}{q}$ to develop a test for $\mathrm{H}_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{t}$. The distribution of F is assumed to approximately follow an F distribution with numerator degrees of freedom q and unknown denominator degrees of freedom ν. Fai and Cornelius (1996) proposed a method for multi-degree-of-freedom tests in unbalanced split-plot designs. As presented by Rencher and Schaalje (2008), this method involves a spectral decomposition of $[\mathbf{L}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{L}']^{-1}$ to yield $\mathbf{P}'[\mathbf{L}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{L}']^{-1}\mathbf{P} = \mathbf{D}$, where $\mathbf{D} = \mathrm{diag}(\lambda_1,\ \lambda_2, \ldots,\ \lambda_\mathrm{m})$ is the diagonal matrix of eigenvalues and $\mathbf{P} = (\mathbf{p}_1,\ \mathbf{p}_2, \ldots,\ \mathbf{p}_\mathrm{m})$ is the orthogonal matrix of normalized eigenvectors of $[\mathbf{L}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{L}']^{-1}$. With this decomposition, $Q = qF$ can be written as a sum of q approximate independent squared t-variables,

$$Q = \sum_{\mathrm{i}=1}^{q} \frac{(\mathbf{p}_{\mathrm{i}}'\mathbf{L}\hat{\boldsymbol{\beta}})^2}{\lambda_{\mathrm{i}}} = \sum_{\mathrm{i}=1}^{q} t_{\nu_{\mathrm{i}}}^2 ,$$

where $\mathbf{p}_{\mathrm{i}}'$ is the i-th eigenvector with the respective eigenvalue λ_{i} and ν_{i} is the approximate degrees of freedom for the i-th single degree of freedom t-test (Rencher and Schaalje, 2008).

The values of ν_{i}'s are computed by repeatedly applying a method for single degrees of freedom contrasts (Giesbrecht and Burns, 1985). Following Satterthwaite's premise (1941), this method assumes that for a single-degree-of-freedom test of $\mathrm{H}_0 : \mathbf{c}'\boldsymbol{\beta} = \mathbf{0}$, where $\mathbf{c}$ is a vector of constants, the test statistic

$$t = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{c}'(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{c}}}$$

approximately follows a t distribution with unknown degrees of freedom of ℓ. It also suggests that ℓ can be approximated as

$$\ell \doteq \frac{2[\mathbf{c}'(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{c}]^2}{\mathrm{var}[\mathbf{c}'(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{c}]} .$$

The denominator of this expression can be estimated using the multivariate delta method proposed by Lehmann (1999). Since a squared t-variable with degrees of freedom ν_{i} is an F-variable with degrees of freedom 1 and ν_{i},

$$E(Q) = E\left(\sum_{\mathrm{i}=1}^{q} \mathrm{t}_{\nu_{\mathrm{i}}}^2\right) = E\left(\sum_{\mathrm{i}=1}^{q} F_{1,\nu_{\mathrm{i}}}\right) = \sum_{\mathrm{i}=1}^{q} [E(F_{1,\nu_{\mathrm{i}}})] = \sum_{\mathrm{i}=1}^{q} \frac{\nu_{\mathrm{i}}}{\nu_{\mathrm{i}} - 2} .$$

Furthermore, ν can be found by the relationship $F = q^{-1}Q$, which approximately distributes as $F_{q,\nu}$. $E(F) = \dfrac{E(Q)}{q} = \dfrac{\nu}{\nu - 2}$ and the Satterthwaite DDF can be obtained as

$$\nu = \frac{2E(Q)}{E(Q)-q} = 2\left(\sum_{i=1}^{q}\frac{\nu_i}{\nu_i - 2}\right)\Bigg/\left[\left(\sum_{i=1}^{q}\frac{\nu_i}{\nu_i - 2}\right) - q\right].$$

2.5 Kenward-Roger Method

Similar to the Satterthwaite method, the Kenward-Roger (K-R) method is to approximate a small-sample F distribution. First of all, this method implements adjustments for two sources of bias in $[\mathbf{L}(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-}\mathbf{L}']^{-1}$, an estimator of the covariance matrix of $\mathbf{L}\hat{\boldsymbol{\beta}}$ for small samples. The first source of bias, the variability in $\hat{\boldsymbol{\sigma}}$, is adjusted using an approximation given by Kackar and Harville (1984); the second source of bias, the small sample bias, is adjusted using a method proposed by Kenward and Roger (1997). Both of these adjustments are based on a Taylor series expansion around $\boldsymbol{\sigma}$ (McCulloch et al., 2008). Rencher and Schaalje (2008) specify the form of the adjusted approximate covariance of $\mathbf{L}\hat{\boldsymbol{\beta}}$,

$$\hat{\boldsymbol{\Sigma}}^{*}_{\mathbf{L}\hat{\boldsymbol{\beta}}} = \mathbf{L}\left\{(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-} + 2(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-}\left[\sum_{i=0}^{m}\sum_{j=0}^{m} s_{ij}(\mathbf{G}_{ij} - \mathbf{P}_i\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\mathbf{P}_j)\right](\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-}\right\}\mathbf{L}',$$

where s_{ij} is the (i, j)-th element of $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\sigma}}}$, $\mathbf{G}_{ij} = \mathbf{X}'\dfrac{\partial\hat{\boldsymbol{\Sigma}}^{-1}}{\partial\sigma_i^2}\hat{\boldsymbol{\Sigma}}\dfrac{\partial\hat{\boldsymbol{\Sigma}}}{\partial\sigma_i^2}\mathbf{X}$ and $\mathbf{P}_i = \mathbf{X}'\dfrac{\partial\hat{\boldsymbol{\Sigma}}^{-1}}{\partial\sigma_i^2}\mathbf{X}$.

The Kenward-Roger method then assumes that for the test $\mathrm{H}_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{t}$, the test statistic,

$$F^{*} = \delta F_{\mathrm{KR}} = \frac{\delta}{q}(\mathbf{L}\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\Sigma}}^{*}_{\mathbf{L}\hat{\boldsymbol{\beta}}}(\mathbf{L}\hat{\boldsymbol{\beta}}),$$

approximately distributes as an F distribution with two adjustable constants, a scale factor δ and the denominator degrees of freedom ν. Based on a second-order Taylor series expansion of $\hat{\boldsymbol{\Sigma}}^{*}_{\mathbf{L}\hat{\boldsymbol{\beta}}}$ around $\boldsymbol{\sigma}$ and the conditional expectation relationships $\mathrm{E}(F_{\mathrm{KR}})$ and $\mathrm{var}(F_{\mathrm{KR}})$ are yielded approximately (Schaalje et al., 2002). By further equating $\mathrm{E}(F_{\mathrm{KR}})$ and $\mathrm{var}(F_{\mathrm{KR}})$ to the mean and variance of an *F* distribution to solve for δ and ν, the results (Rencher and Schaalje, 2008) can be obtained as

$$\nu = 4 + \frac{q+2}{q\gamma - 1}$$

and

$$\delta = \frac{\nu}{\mathrm{E}(F_{\mathrm{KR}})(\nu - 2)},$$

where

$$\gamma = \frac{\mathrm{var}(F_{\mathrm{KR}})}{2\mathrm{E}(F_{\mathrm{KR}})^2}.$$

3 Simulation

3.1 Settings

To see how each method of the denominator degrees of freedom approximations handles a relatively complicated mixed model, in the simulation we introduce a model of time points with random intercept and random slope,

$$Y_{ij} = \gamma_0 + \gamma_1 t_{ij} + \gamma_2 t_{ij}^2 + \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij},$$

$$i = 1,\ 2, \ldots, n;\ j = 1,\ 2, \ldots, m,$$

where Y_{ij} is the j-th observation for the i-th factor level; γ_0 is the intercept for the fixed effects; γ_1 is the slope for the time fixed effects; γ_2 is the slope for the time-square (timesq) fixed effects; β_{0i} is the random intercept that distributes as $N(0,\ \sigma_{b_0}^2)$; β_{1i} is the random slope that distributes as $N(0,\ \sigma_{b_1}^2)$. ε_{ij} is the noise identically independently distributed as $N(0,\ \sigma^2)$.

Our simulations are conducted through a combination of code in R and SAS (Appendix A, B, C) using the REML approach and the unstructured type of covariance components. We have three types of setting variables as follows:

Table 1

Simulation Setting Variables

No.	Type	Setting Variables
1	Sample structure $(n \times m)$	$(20\times3), (20\times6), (10\times6)$
2	Slopes for time and timesq $(\gamma_1,\ \gamma_2)$	$(0, 0), (0.05, 0.05), (0.1, 0.1), (0.5, 0.5)$
3	Proportion of missing data	0% (balanced), 5%, 10%, 20%

There are several steps of how we conduct the simulation: first of all, for each type of the sample structure, we modify Adriaenssens' R code (2015) to generate 500 sets of random variables for each setting of slopes for time and timesq. We then use the Sample command in R to randomly omit certain amount of values in the dataset and fulfill the requirements of the missing-data proportion. For each dataset generated from each of the 48 setting-variable combinations in total, we use the Proc Mixed statement in SAS to run the mixed model by five DDF approximation methods (residual, containment, between-within, Satterthwaite, and Kenward-Roger). Finally, we export the results for each method and plot F-statistic quantile-quantile (Q-Q) plots, DDF boxplots, and p-value histograms to get insights on each method's performance.

3.2 Results

This section is divided into five parts with respect to the five methods of denominator degrees of freedom approximations. In each part, we first analyze the F-statistic Q-Q plots with $(\gamma_1, \gamma_2) = (0, 0)$. Then we evaluate the DDF boxplots with different settings of sample structure, slopes, and missing-data proportion. Last, we look into the p-value histograms to conclude with the change of power across different designs. Note that we only present figures for results that are representative of their category or particularly deviant from the common pattern of their category.

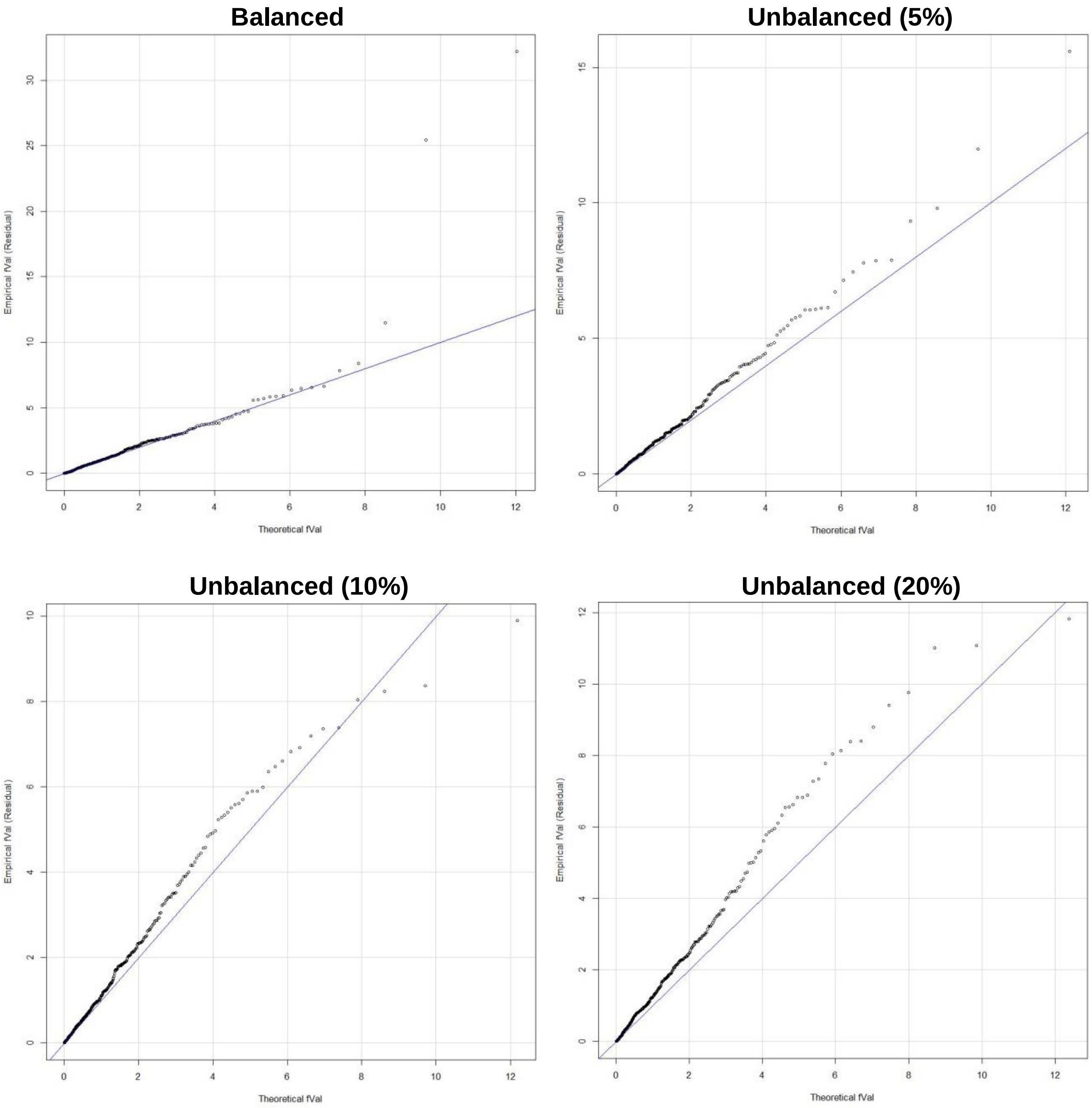
3.2.1 Residual Method

3.2.1.1 F-Statistic Q-Q Plots. Given that when the slopes of time and timesq equal to (0, 0) in the model, we have a central F distribution for the test statistic. It is assumed that in the case with $(\gamma_1, \gamma_2) = (0, 0)$, the empirical F-statistics computed by the DDF approximation methods should be very similar to the theoretical F-statistics.

Looking into Figure 1 below, we observe that this is very likely the case for balanced designs with only a few outliers out of 500 data. However, as the proportion of missing data increases, the deviance between the empirical F-statistics and the theoretical F-statistics goes up.

Figure 1

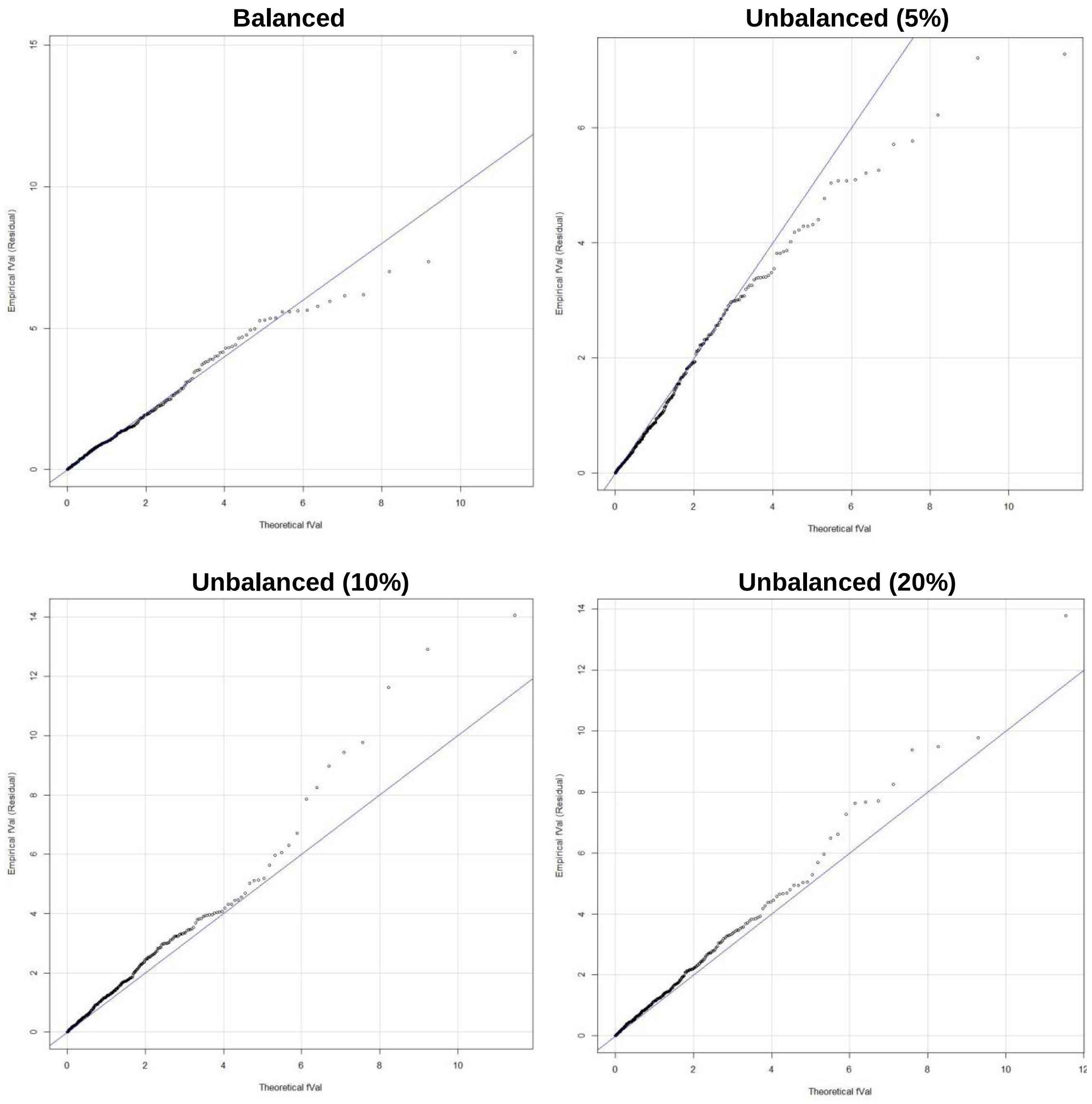
Residual F-Statistic Q-Q Plots Highlighted with 20×3 Time Effects



This positive relationship between the imbalance proportion and the F-Statistic deviance also applies to the case of 20×3 timesq effects. Further looking into Figure 2 below, we observe that a different pattern is present when we have more time points (six compared to three) nested within each factor levels.

Figure 2

Residual F-Statistic Q-Q Plots Highlighted with 20×6 Time Effects

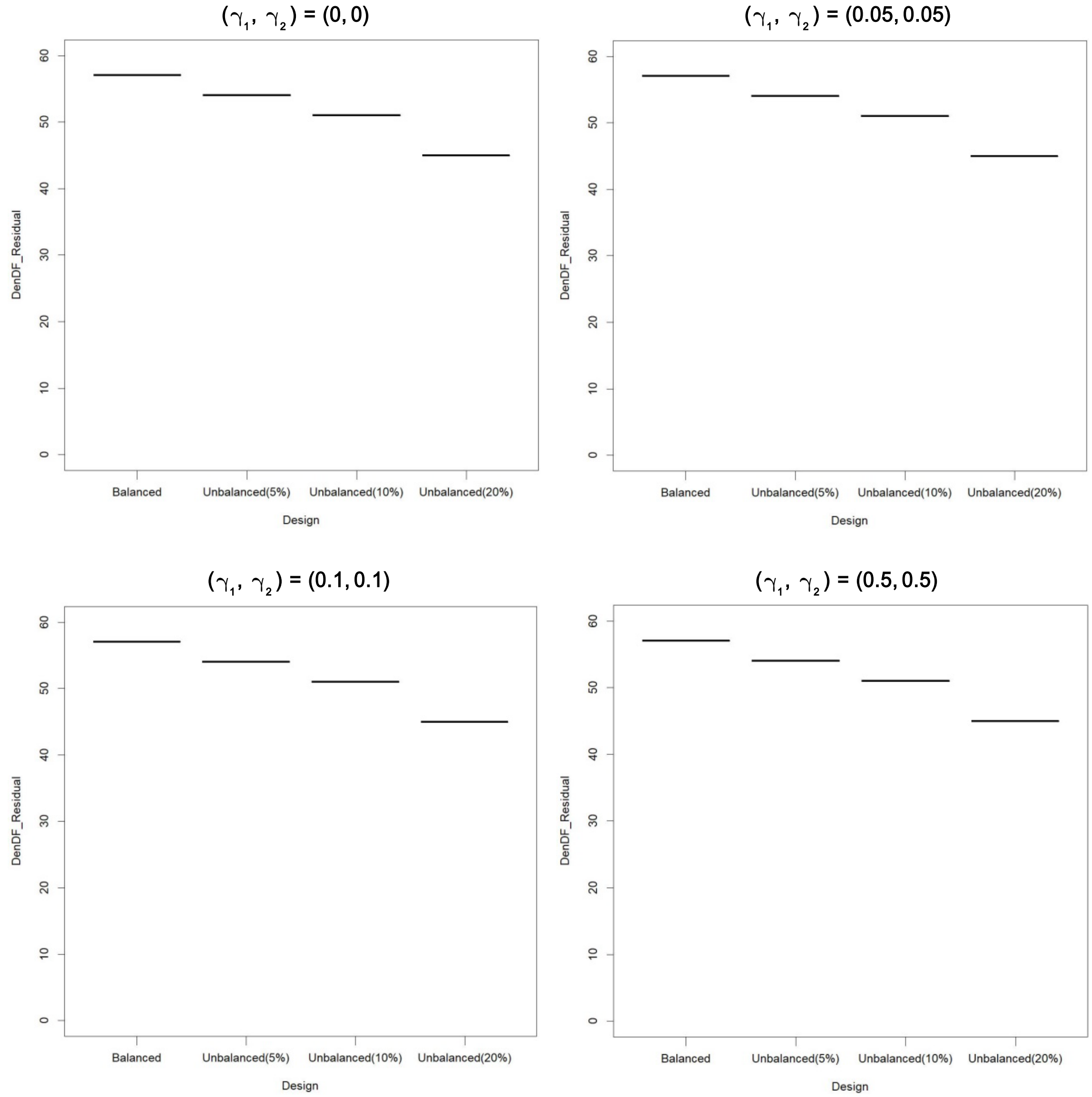


The difference between the empirical F-Statistic and the theoretical F-Statistic becomes smaller for the datasets with larger proportion of missing values. This change of pattern is present for 20×6 timesq effects and 10×6 time and timesq effects as well.

3.2.1.2 DDF Boxplots. The DDF distribution is robust across different slopes.

Figure 3

Residual DDF Boxplots Highlighted with 20×3 Time Effects

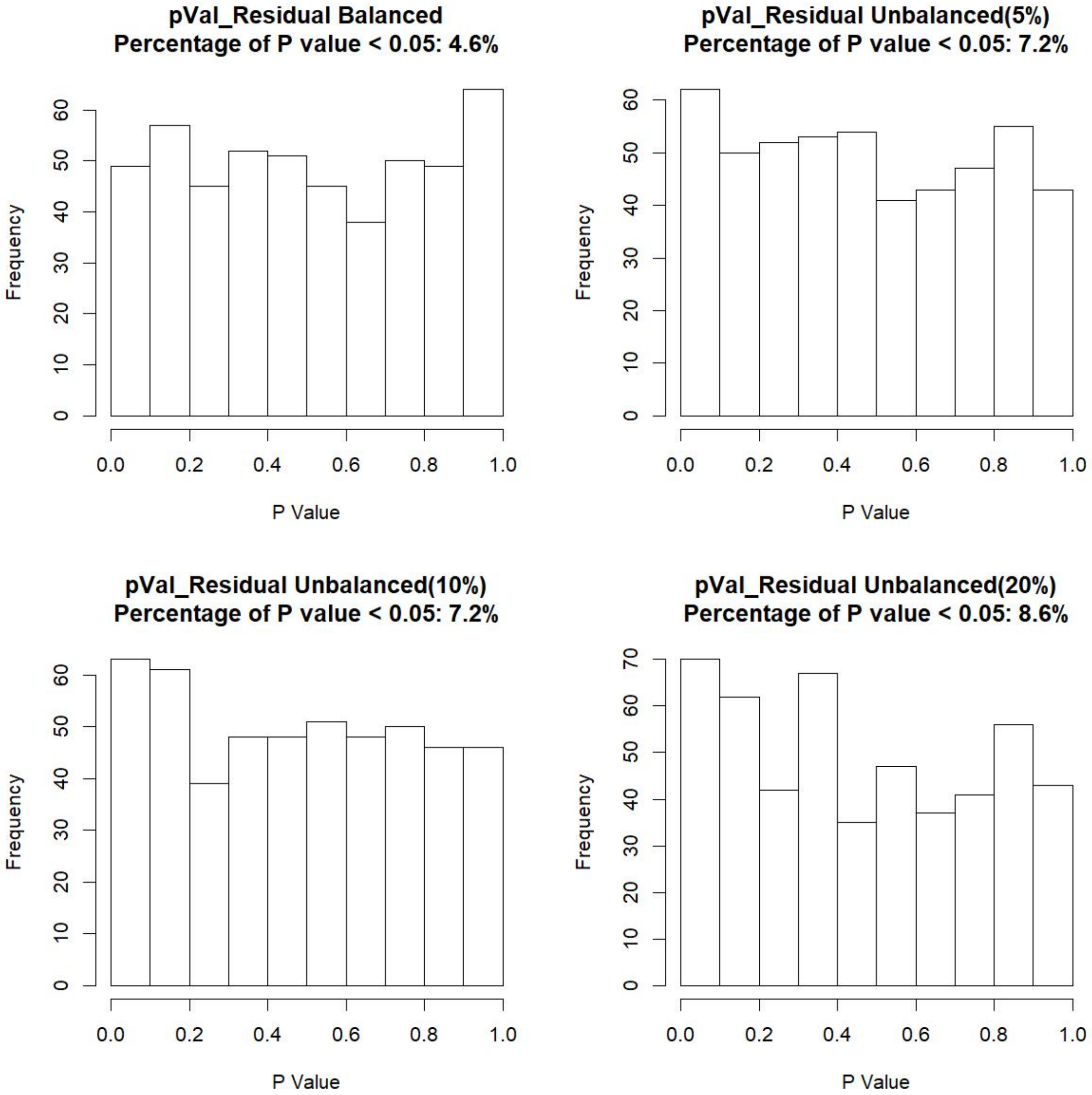


As showed in Figure 3 on the previous page, the distribution of DDF is very concentrated with no observable extreme values. As the proportion of missing data increases, the mean of DDF decreases. Similar trend can be found for 20×3 timesq effects, 20×6 time and timesq effects, and 10×6 time and timesq effects.

3.2.1.3 P-Value Histograms. First, the case with $(\gamma_1, \gamma_2) = (0, 0)$ is as follows:

Figure 4

Residual P-Value Histograms for 20×3 (0, 0) Time Effects



With slopes equal to zero, the percentage of p-values less than $\alpha = 0.05$ should be around 5%. In Figure 4 on the previous page, the mean percentage for the four designs with different proportion of missing data is 6.9%. Since in each trial of the hypothesis test the result is either rejecting or not rejecting the null hypothesis, we apply the normal approximation of binomial distribution to obtain a confidence interval for the percentage:

$$0.05 \pm 1.96\sqrt{\frac{0.05 \times 0.95}{500}} = (3.09\%,\ 6.91\%) .$$

Since the mean 6.9% is within the confidence interval, we conclude that the residual method performs well for 20×3 time effects with $(\gamma_1,\ \gamma_2) = (0, 0)$. Further computing the means for the rest of the designs, we get a full table of mean percentages as follows:

Table 2

Residual Mean Percentage for (γ_1, γ_2) = (0, 0)

No.	Sample Structure	Effect Type	Mean Percentage (%)
1	20×3	Time	6.9
2	20×3	Timesq	7.1*
3	20×6	Time	5.3
4	20×6	Timesq	5.8
5	10×6	Time	5.1
6	10×6	Timesq	4.7

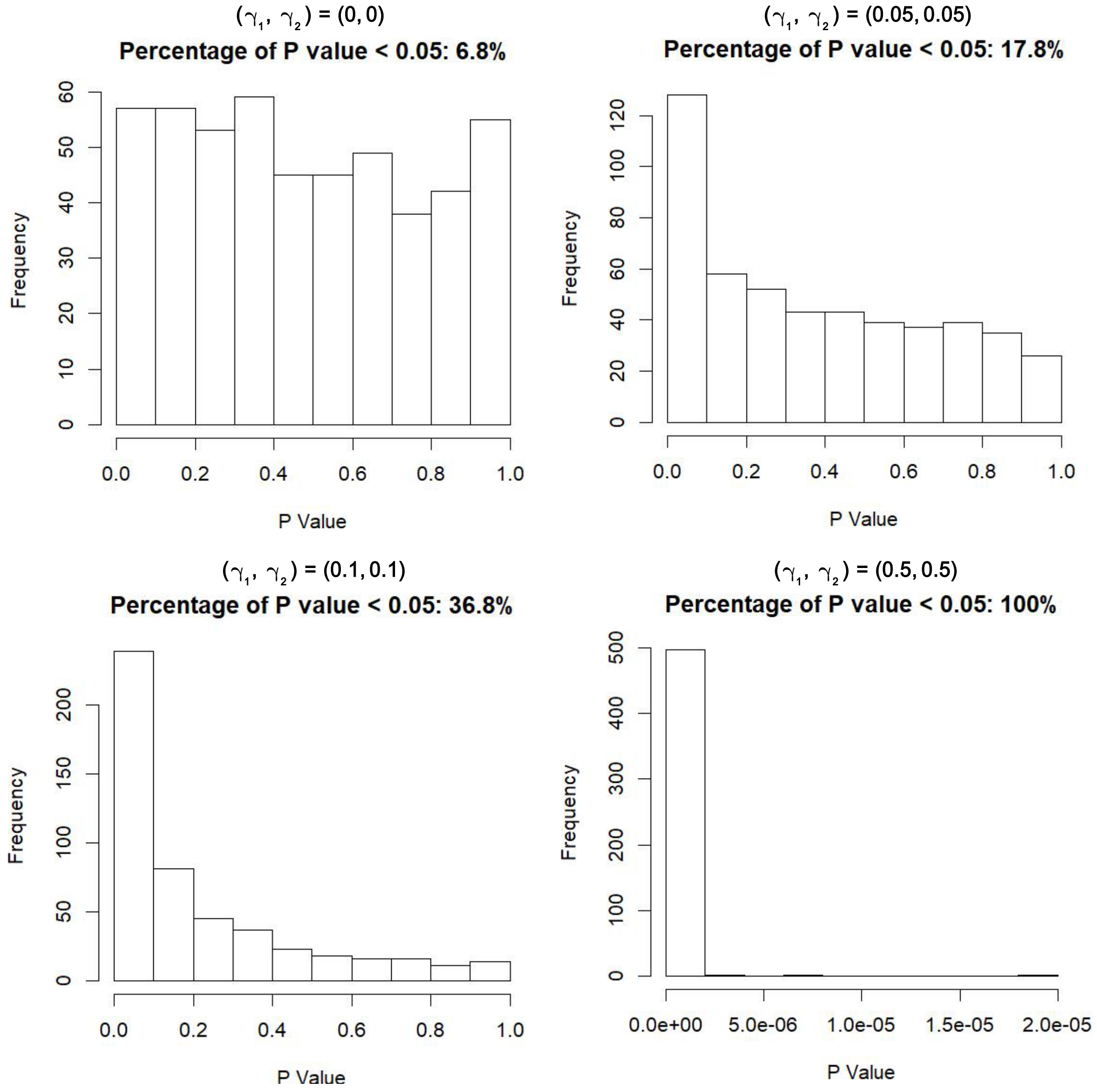
*beyond the confidence interval of (3.09%, 6.91%)

Note there is one mean falling beyond the confidence interval. The sample structure of 20×3 gives the greatest mean percentages for both time and timesq effects among all.

Furthermore, we look into the trend across different settings of slopes. It is reasonable that the higher slopes we set for the effects, the more power we should get in terms of rejecting the null hypothesis. Figure 5 below shows a typical trend of power across different slopes. We also observe another trend that under the same setting of slopes, the power for time effects tends to be less than the one for timesq effects.

Figure 5

Residual P-Value Histograms for 20×3 Unbalanced (5%)Timesq Effects

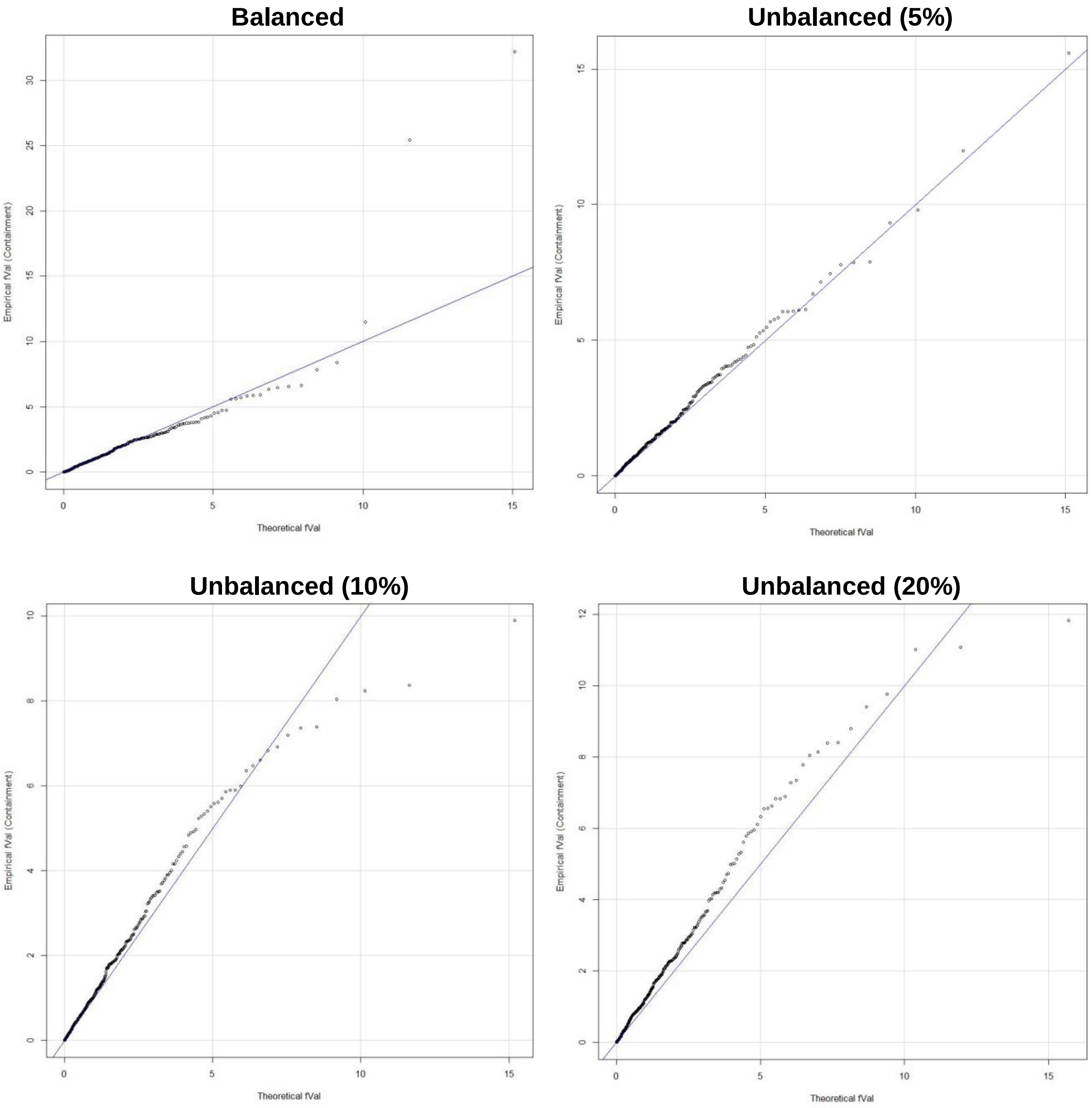


3.2.2 Containment Method

3.2.2.1 F-Statistic Q-Q Plots. Looking into the results for the containment method, we observe a trend similar to the residual method: the more proportion of missing data, the larger deviance between empirical and theoretical F-Statistics (Figure 6).

Figure 6

Containment F-Statistic Q-Q Plots Highlighted with 20×3 Time Effects



Balanced

Unbalanced (5%)

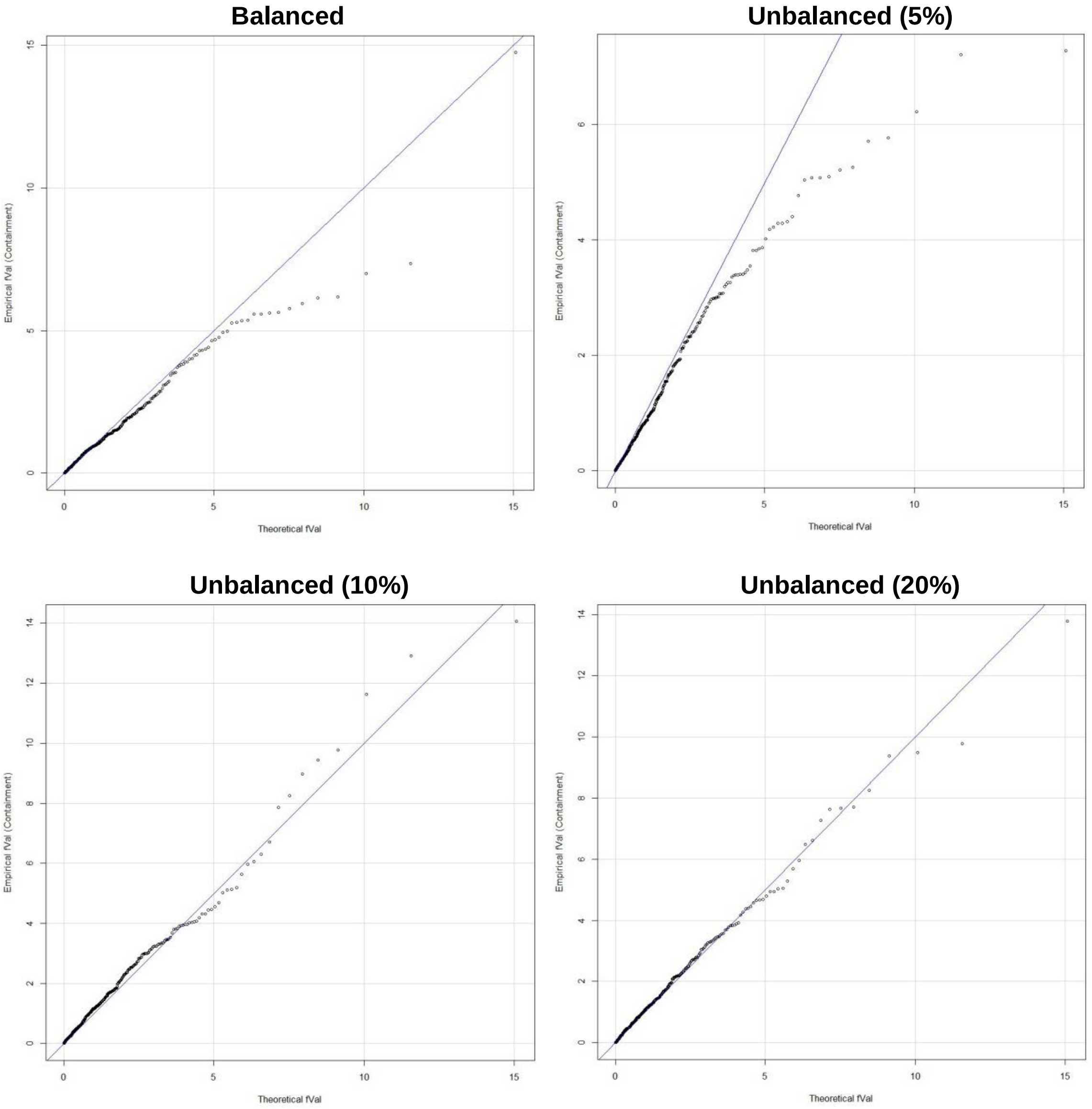
Unbalanced (10%)

Unbalanced (20%)

The exception occurs when we increase the number of time points nested within each of the factor levels from three to six. As presented in Figure 7 below, the F-Statistic deviance with the largest 20% of missing data is relatively small compared to others.

Figure 7

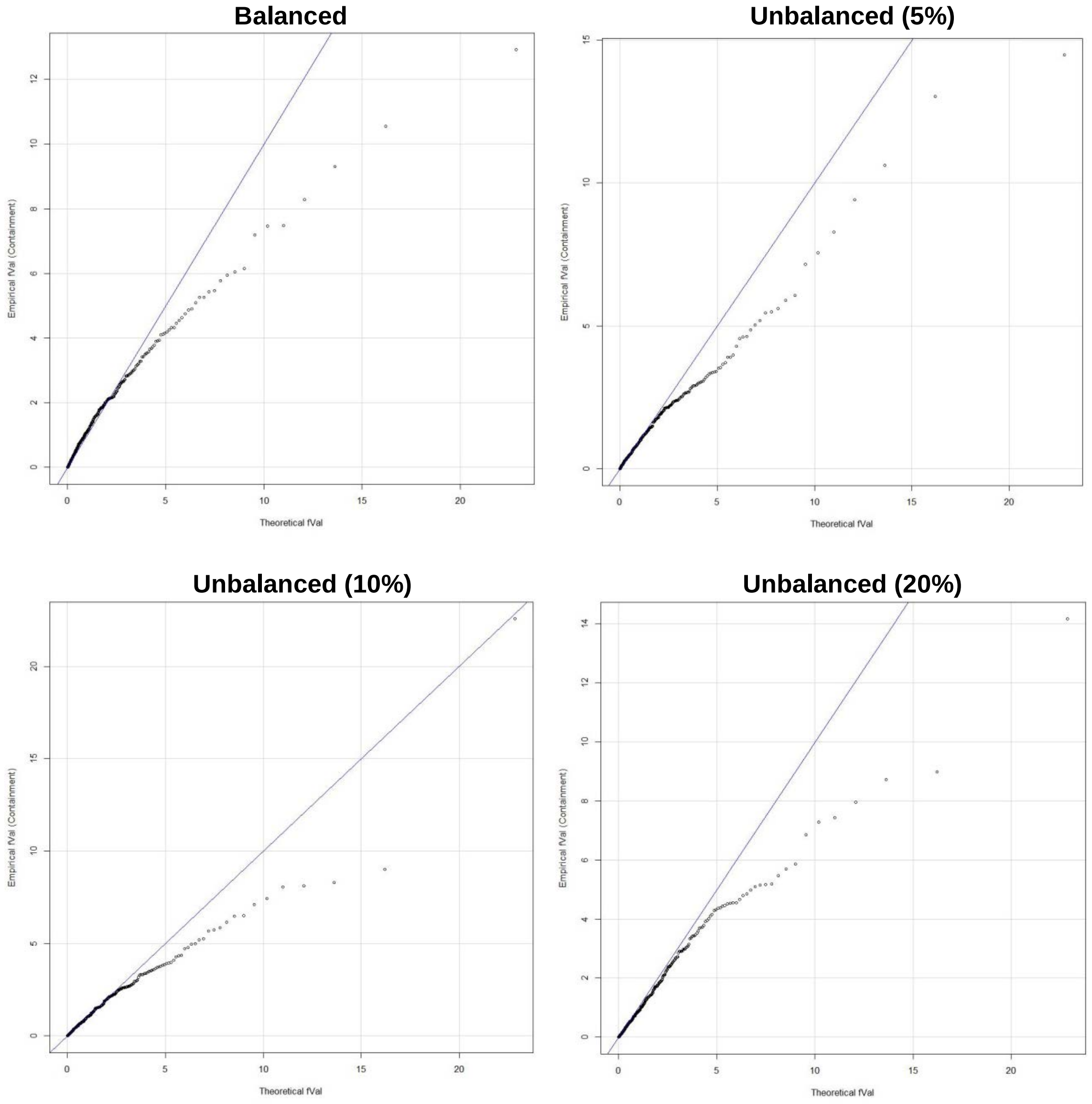
Containment F-Statistic Q-Q Plots Highlighted with 20×6 Time Effects



However, the average F-Statistic deviance becomes larger as we further decrease the number of factor levels from 20 to 10 (Figure 8). Even in the case with balanced data, there is a significant deviance between the empirical and the theoretical F-Statistics.

Figure 8

Containment F-Statistic Q-Q Plots Highlighted with 10 × 6 Time Effects



Balanced

Unbalanced (5%)

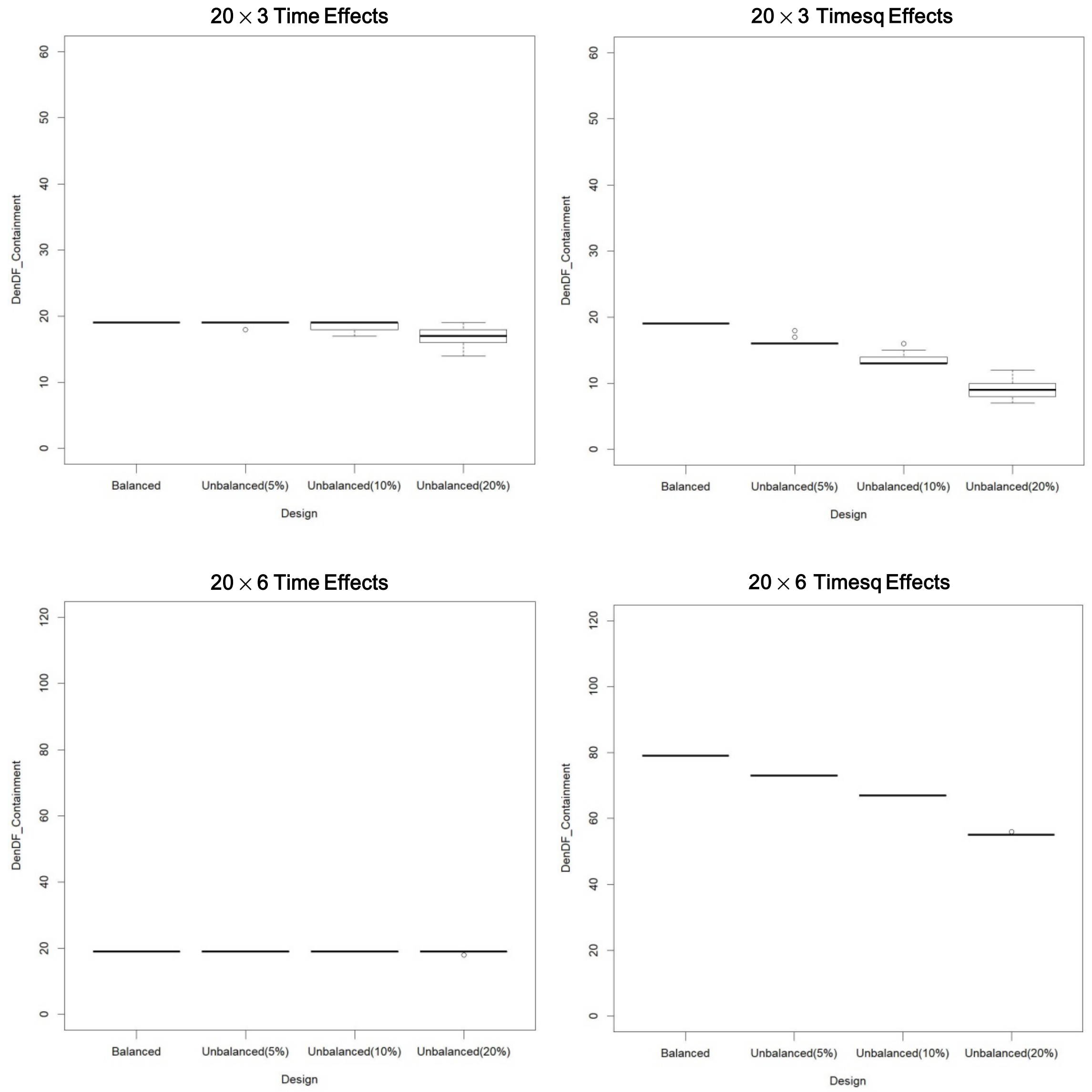
Unbalanced (10%)

Unbalanced (20%)

3.2.2.2 DDF Boxplots. There are two kinds of patterns in the DDF distribution. First, for the time effects with a sample structure of 20×3 (upper left in Figure 9), there is a small deviance among means across different missing-data proportion. Yet the range increases as the proportion goes up. For the timesq effects (upper right in Figure 9), the mean decreases as the proportion increases.

Figure 9

Containment DDF Boxplots Highlighted with $(\gamma_1, \gamma_2) = (0.05, 0.05)$



Second, for the time effects with a sample structure of 20×6 (lower left in Figure 9 on the previous page), the DDF distribution is very concentrated and the mean is almost the same across different proportion of missing values. This pattern is also present for the time effects with a sample structure of 10×6. Moreover, for the timesq effects with a sample structure of 20×6 (lower right in Figure 9); the DDF distribution is still concentrated, but the mean decreases as the missing-data proportion increases. This pattern is present for the timesq effects with a sample structure of 10×6 as well.

3.2.2.3 P-Value Histograms. First of all, we obtain the mean percentages of p-values less than $\alpha = 0.05$ across different missing-data proportion as follows:

Table 3

Containment Mean Percentage for (γ_1, γ_2) = (0, 0)

No.	Sample Structure	Effect Type	Mean Percentage (%)
1	20×3	Time	5.7
2	20×3	Timesq	5.2
3	20×6	Time	4.0
4	20×6	Timesq	5.7
5	10×6	Time	2.8
6	10×6	Timesq	4.5

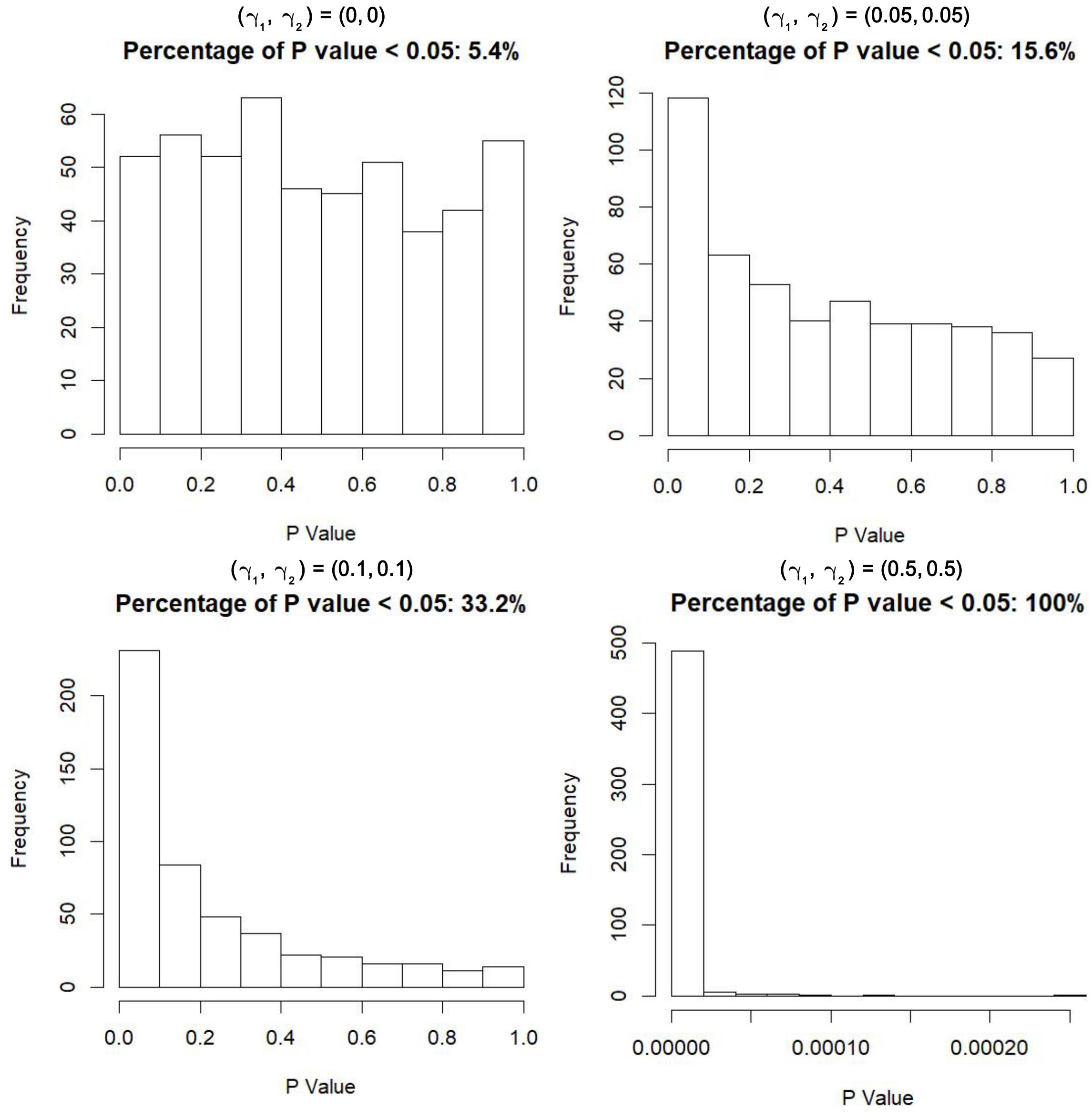
*beyond the confidence interval of (3.09%, 6.91%)

Note there is no mean percentage falling beyond the confidence interval. Compared to the previous mean percentage of 7.1% for the time effects with a sample structure of 20×3, this result suggests that the containment method performs better than the residual method in terms of percentages of p-values less than 0.05 with $(\gamma_1, \gamma_2) = (0, 0)$.

Further looking into the p-values across different settings of slopes, we observe a trend similar to the residual method. Figure 10 below shows an increasing trend of power across different slopes. The power for time effects also tends to be less than the one for timesq effects with the same setting of slopes.

Figure 10

Containment P-Value Histograms for 20×3 Unbalanced (5%)Timesq Effects

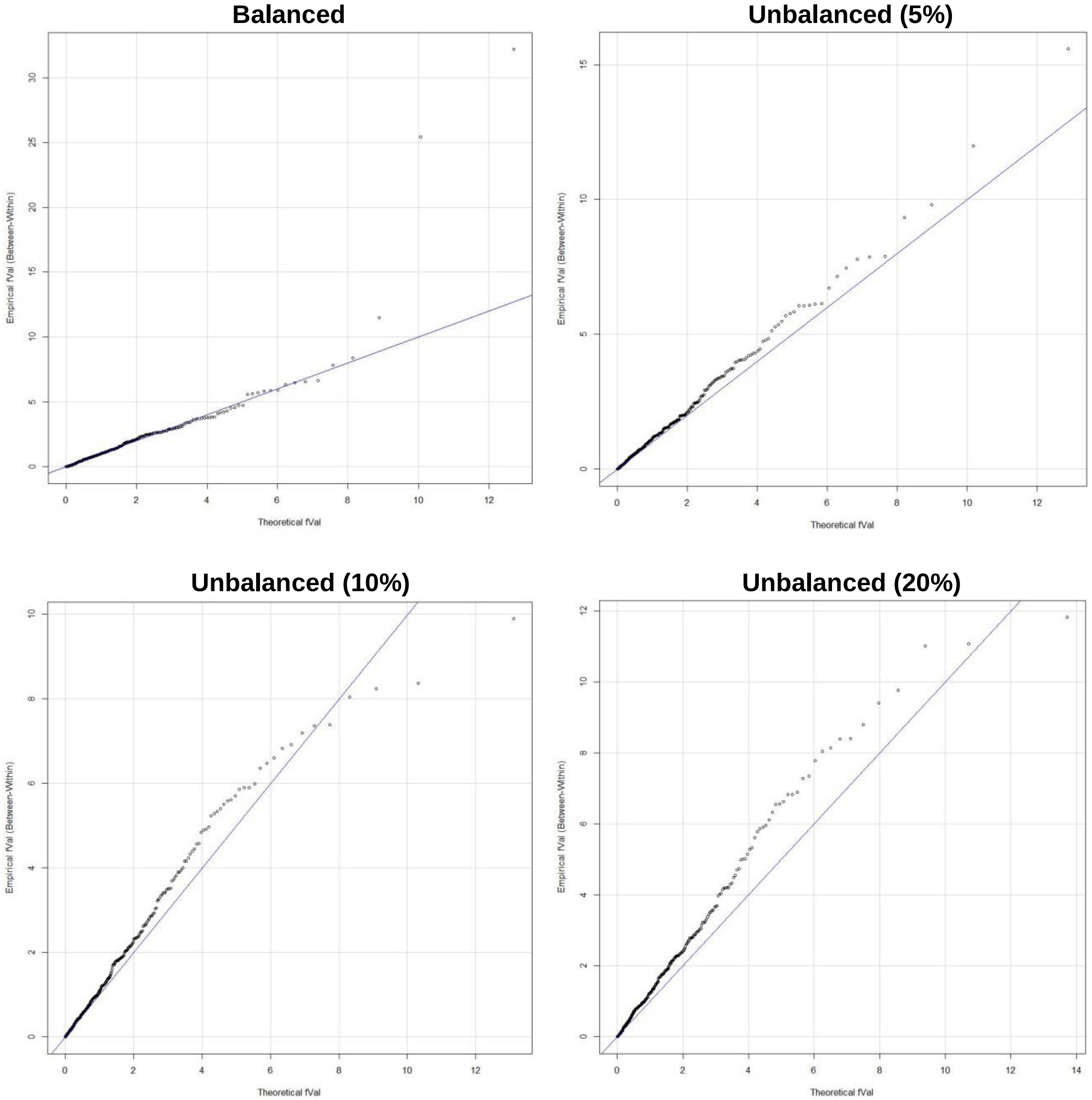


3.2.3 Between-Within Method

3.2.3.1 DDF F-Statistic Plots. First, we observe that the previous positive relationship between the missing-data proportion and the F-Statistic deviance is present for 20×3 time and timesq effects when the slopes for time and timesq equal to (0, 0), as illustrated in Figure 11 below.

Figure 11

B-W F-Statistic Plots Highlighted with 20 × 3 Time Effects



Balanced

Unbalanced (5%)

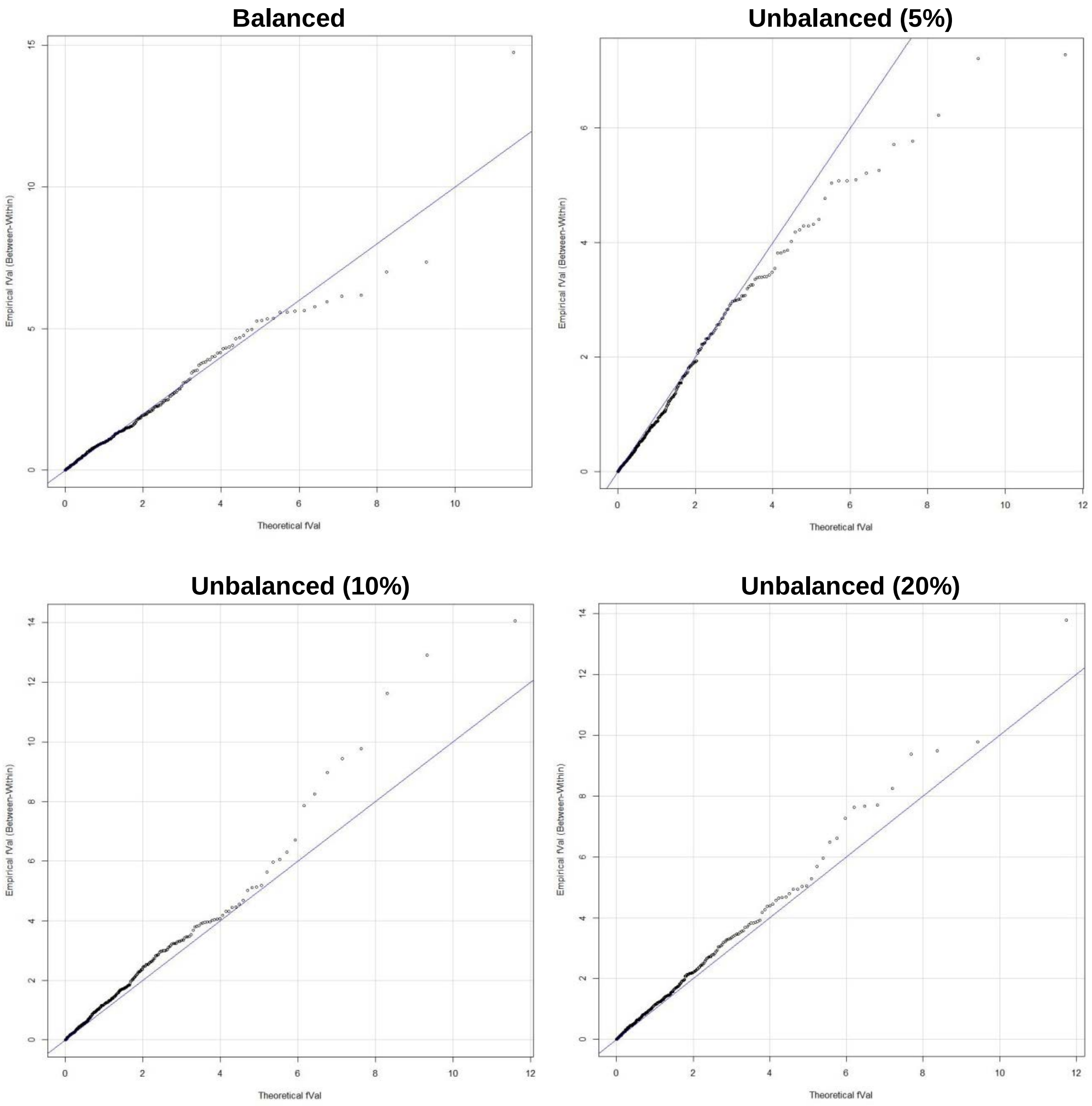
Unbalanced (10%)

Unbalanced (20%)

Nonetheless, when we increase the number of time points nested within each factor level from three to six, the deviance between the empirical and the theoretical F-Statistics becomes smaller as the proportion of missing values increases (Figure 12). This pattern also applies to the cases of 20×6 timesq effects and 10×6 time and timesq effects.

Figure 12

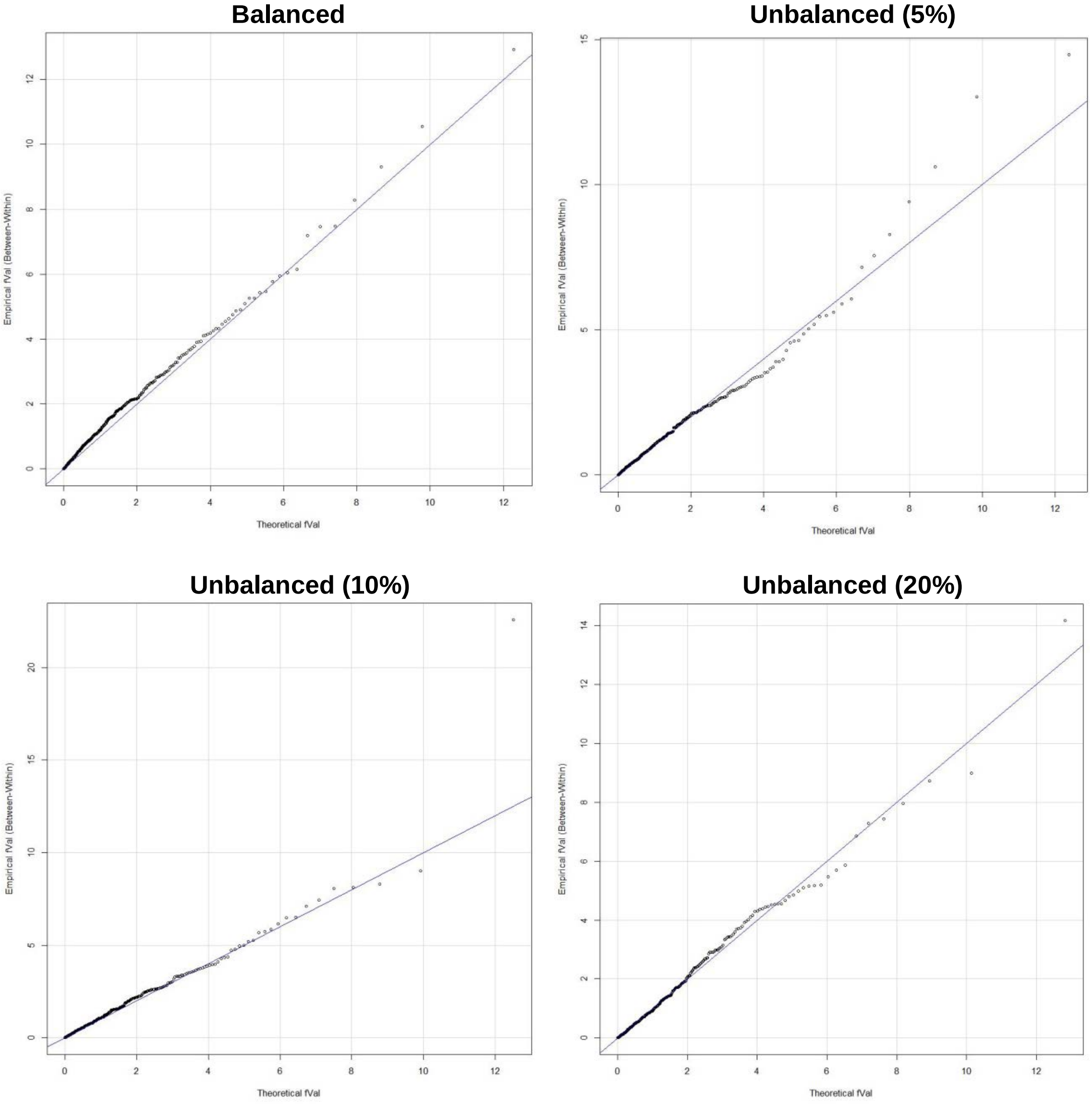
B-W F-Statistic Plots Highlighted with 20×6 Time Effects



When we further decrease the number of factor levels, we observe that the F-Statistic deviance growing along with the missing-data proportion becomes smaller (Figure 13) compared to the previous cases with the sample structure of 20×3 and 20×6.

Figure 13

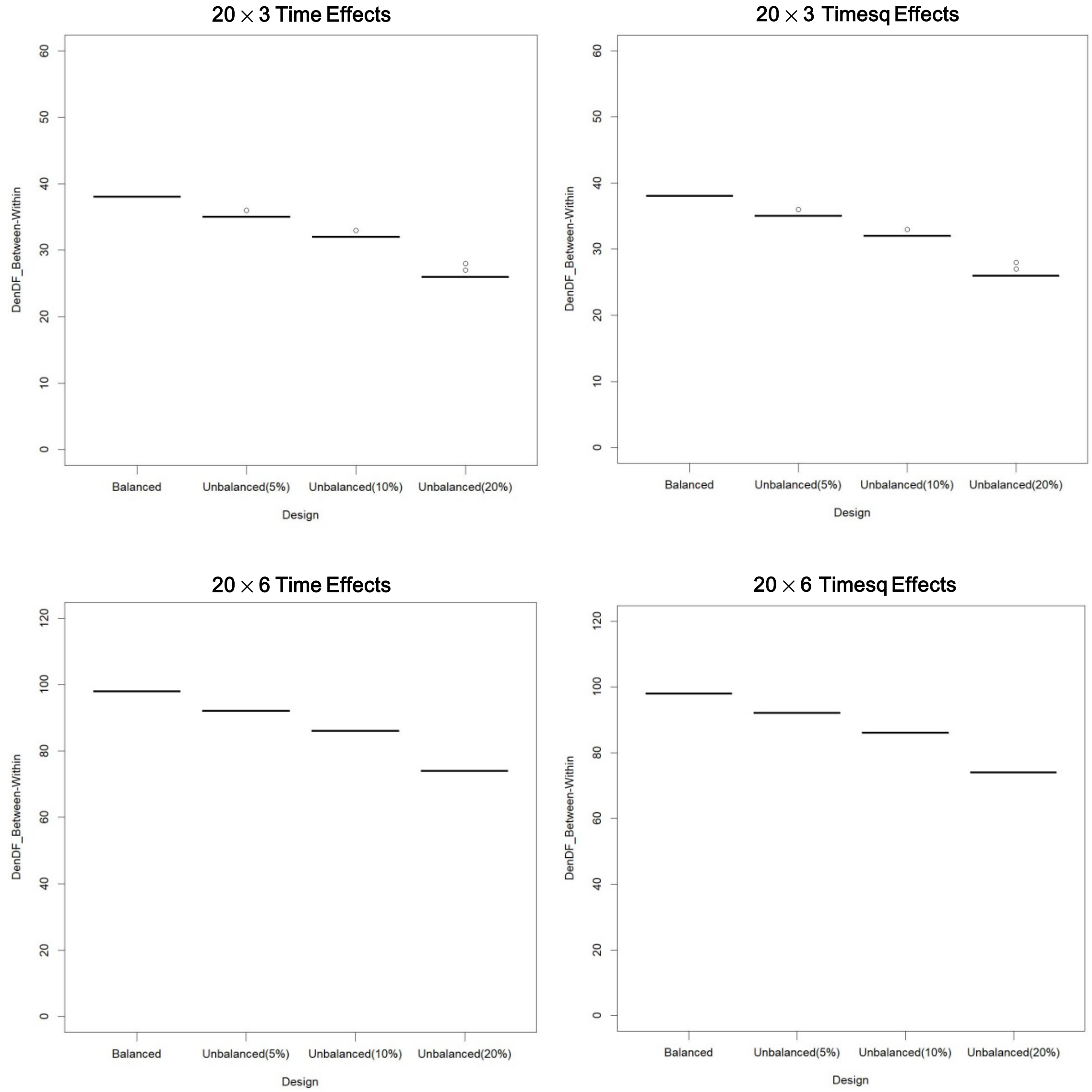
B-W F-Statistic Plots Highlighted with 10×6 Time Effects



3.2.3.2 DDF Boxplots. The DDF distribution with respect for each type of missing-data proportion is robust regardless of the kind of effects. There is a universal trend for the B-W method that the DDF mean decreases as the missing-data proportion increases. The small amount of outliers presented with the sample structure of 20×3 in Figure 14 below does not exist for the cases of 20×6 and 10×6 sample structure.

Figure 14

B-W DDF Boxplots Highlighted with (γ_1, γ_2) = (0.05, 0.05)



20×3 Time Effects

20×3 Timesq Effects

20×6 Time Effects

20×6 Timesq Effects

3.2.3.3 P-Value Histograms. First, we obtain the mean percentages of p-values less than $\alpha = 0.05$ across different missing-data proportion:

Table 4

B-W Mean Percentage for (γ_1, γ_2) = (0, 0)

No.	Sample Structure	Effect Type	Mean Percentage (%)
1	20×3	Time	6.4
2	20×3	Timesq	6.8
3	20×6	Time	5.3
4	20×6	Timesq	5.8
5	10×6	Time	5.0
6	10×6	Timesq	4.6

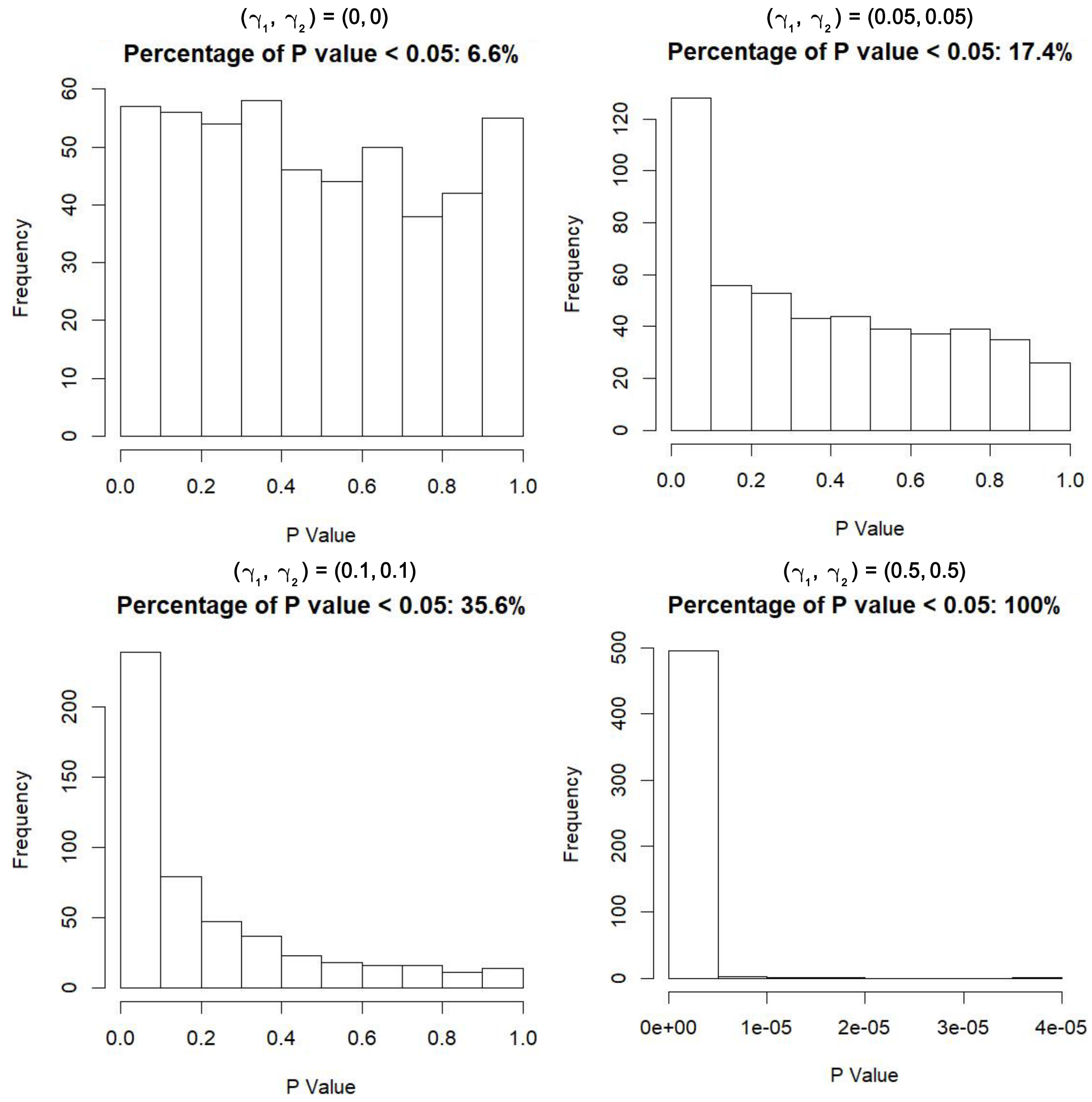
*beyond the confidence interval of (3.09%, 6.91%)

Note there is no mean percentage falling beyond the confidence interval. This result suggests that the B-W method performs better than the residual method in terms of percentages of p-values less than 0.05 with $(\gamma_1, \gamma_2) = (0, 0)$. Nevertheless, the majority of mean percentages here is greater than the ones produced by the containment method.

A typical trend is shown in Figure 15 on the next page: As the slopes for time and timesq effects increase, the power also increases. This trend is present in all of the designs regardless of the type of effects, sample structure or missing-data proportion. However, the percentage of p-values less than $\alpha = 0.05$ given by the B-W method tends to be greater than the one given by the containment method.

Figure 15

B-W P-Value Histograms for 20×3 Unbalanced (5%) Timesq Effects



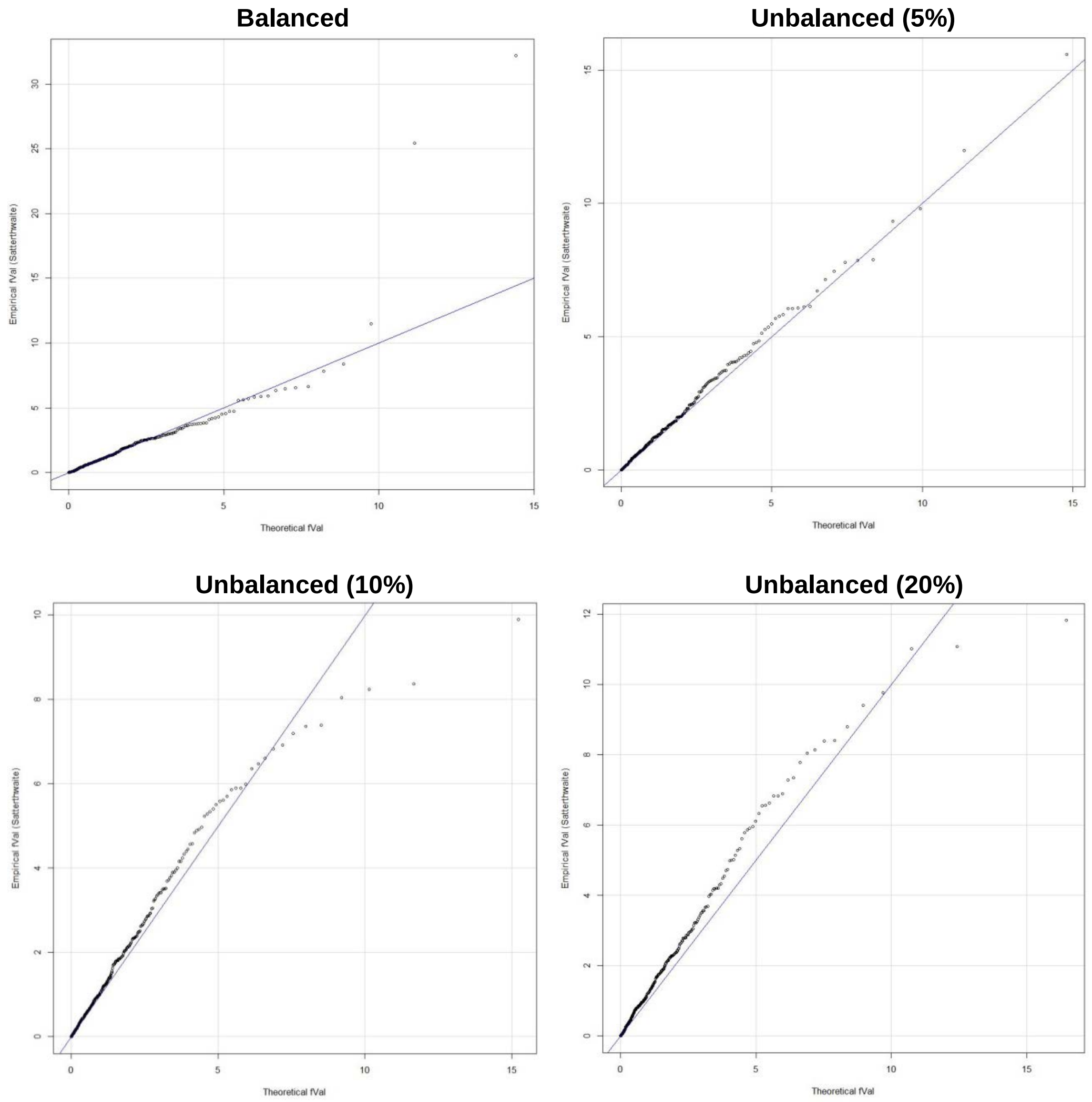
3.2.4 Satterthwaite Method

3.2.4.1 F-Statistic Q-Q Plots. First of all, we observe a positive relationship between the F-Statistic deviance and the missing-data proportion. This pattern occurs with the residual, containment and B-W method as well. Further comparing the results in

Figure 6 on p.18 and those in Figure 16 below, we can see that the way data scatter by the containment method is very similar to the way by the Satterthwaite method. This similarity holds for the majority of the F-Statistic Q-Q plots of these two methods. Therefore, it may suggest a similar way these two methods handle data when the slopes for time and timesq are set to be zero.

Figure 16

Satterthwaite F-Statistic Q-Q Plots Highlighted with 20×3 Time Effects



Balanced

Unbalanced (5%)

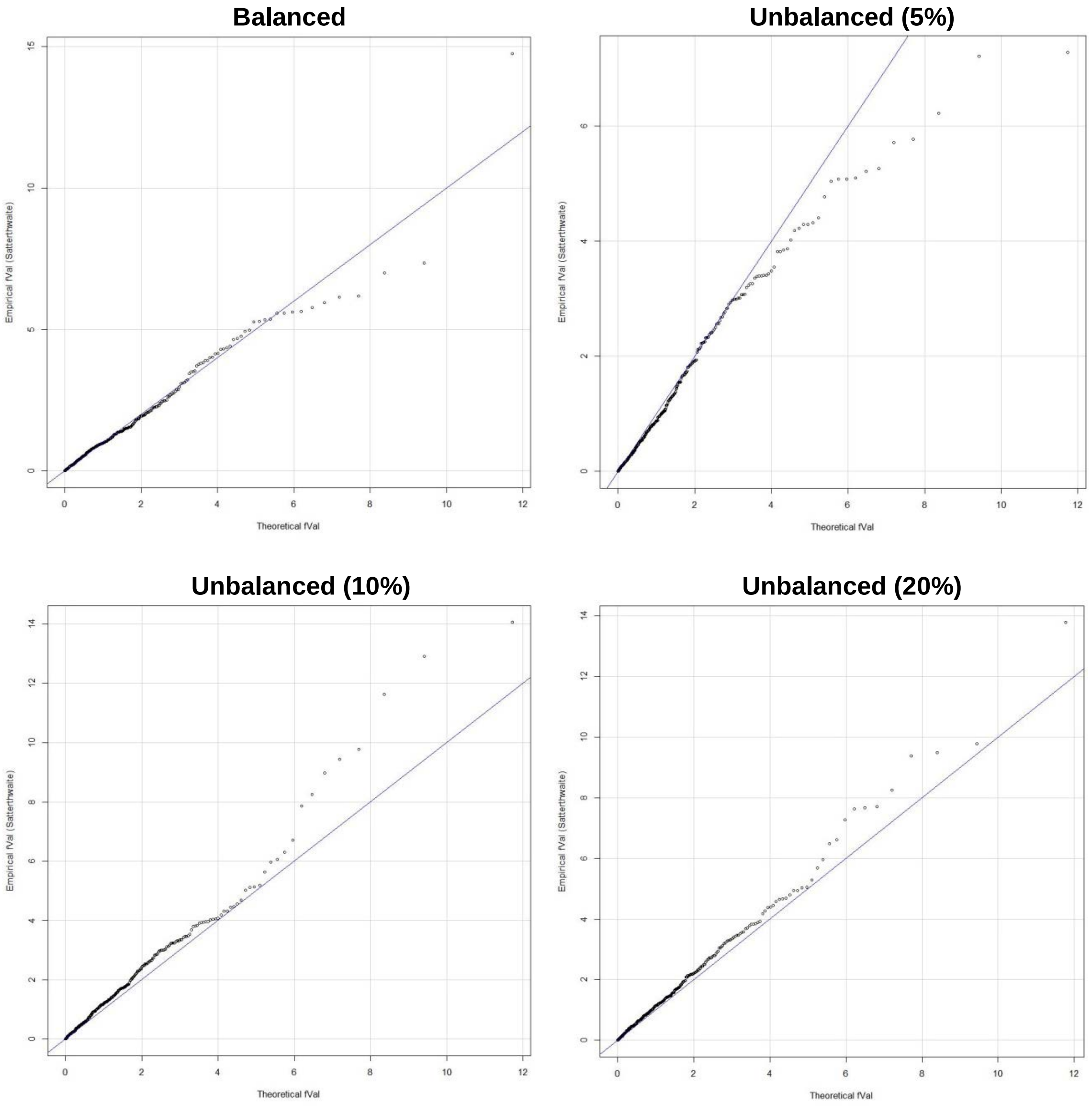
Unbalanced (10%)

Unbalanced (20%)

The F-Statistic deviance for the case with 20% missing values gets smaller as more time points are set to be nested within each factor level. This change, illustrated in Figure 17 below, occurs in the same kind of designs by residual, containment and B-W methods as well.

Figure 17

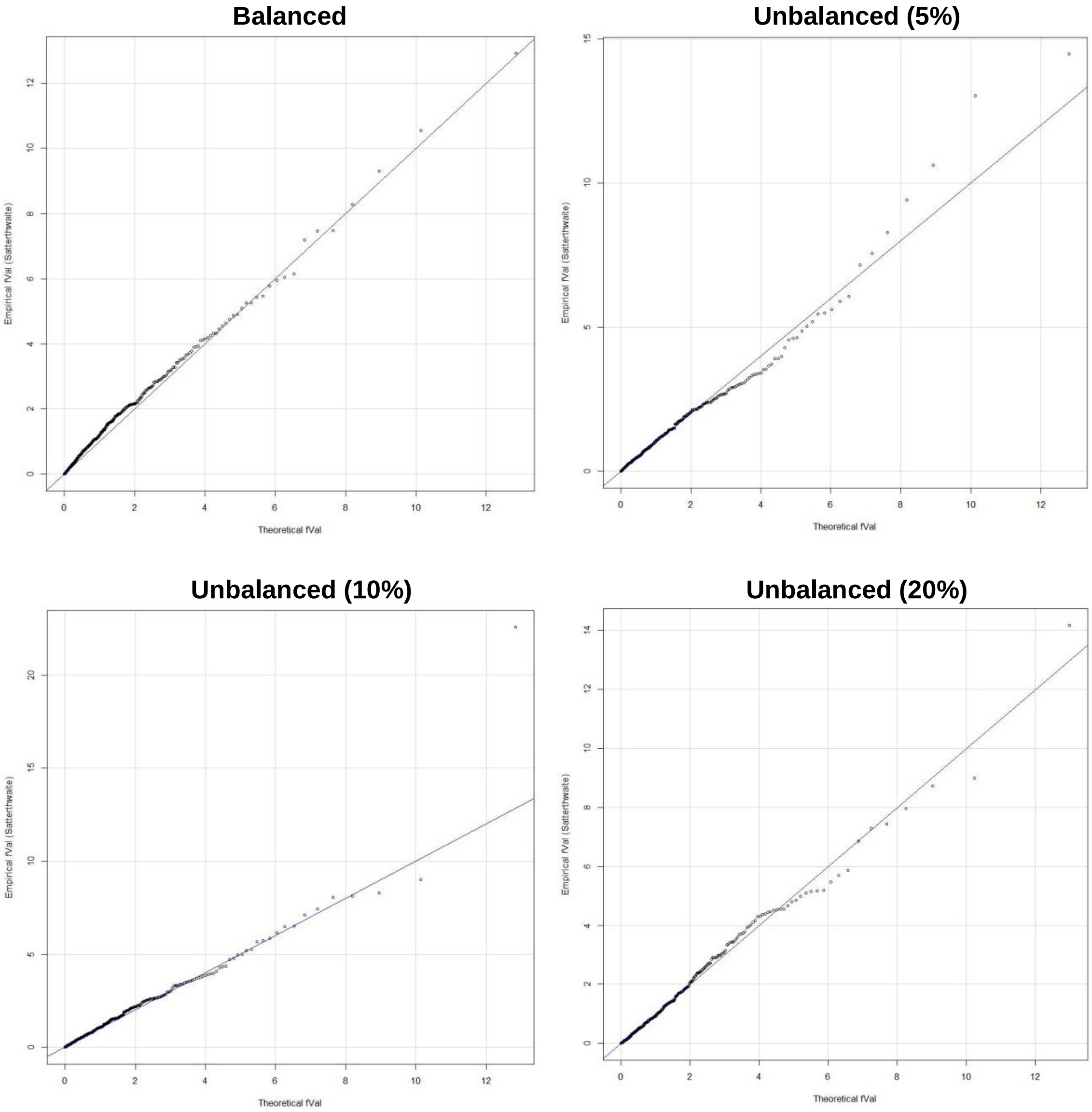
Satterthwaite F-Statistic Q-Q Plots Highlighted with 20×6 Time Effects



The F-Statistic deviance, especially for the extreme values, decreases as we further increase the number of factor levels (Figure 18). Yet compared to Figure 17, some deviance may actually become larger, such as the case with 5% missing data.

Figure 18

Satterthwaite F-Statistic Q-Q Plots Highlighted with 10 × 6 Time Effects



3.2.4.2 DDF Boxplots. Previously the DDF boxplots presented for the residual, containment and B-W methods often share only two kinds of patterns across different designs. Contrarily for the Satterthwaite method, we discover that with each type of sample structure (20×3, 20×6 and 10×6), the DDF boxplots show different patterns with respect to the type of effects, slopes and missing-data proportion.

First of all, in Figure 19 on p.34, we observe that the DDF distribution for the time effects and the timesq effects is quite similar. The mean DDF is also similar with respect to the slopes and missing-data proportion. The mean decreases as the proportion of missing values increases.

Second, in Figure 20 on p.35, the difference between the DDF distributions for the time effects and the timesq effects is quite distinguishing. On one hand, the distribution for the time effects is relatively scattered around similar means. The range also decreases as the missing-data proportion increases. On the other hand, the distribution for the timesq effects is relatively concentrated. The mean decreases as the missing-data proportion increases.

Finally, in Figure 21 on p.36, the difference between the time effects and the timesq effects is also present. However, compared to the case in Figure 20, the distributions for both time and timesq effects are all more scattered. There are also more outliers. This result may be due to the decrease in the number of factor levels (from 20 to 10). Nonetheless, the means follow the same fashion in Figure 20.

Figure 19

Satterthwaite DDF Boxplots with 20×3 Sample Structure

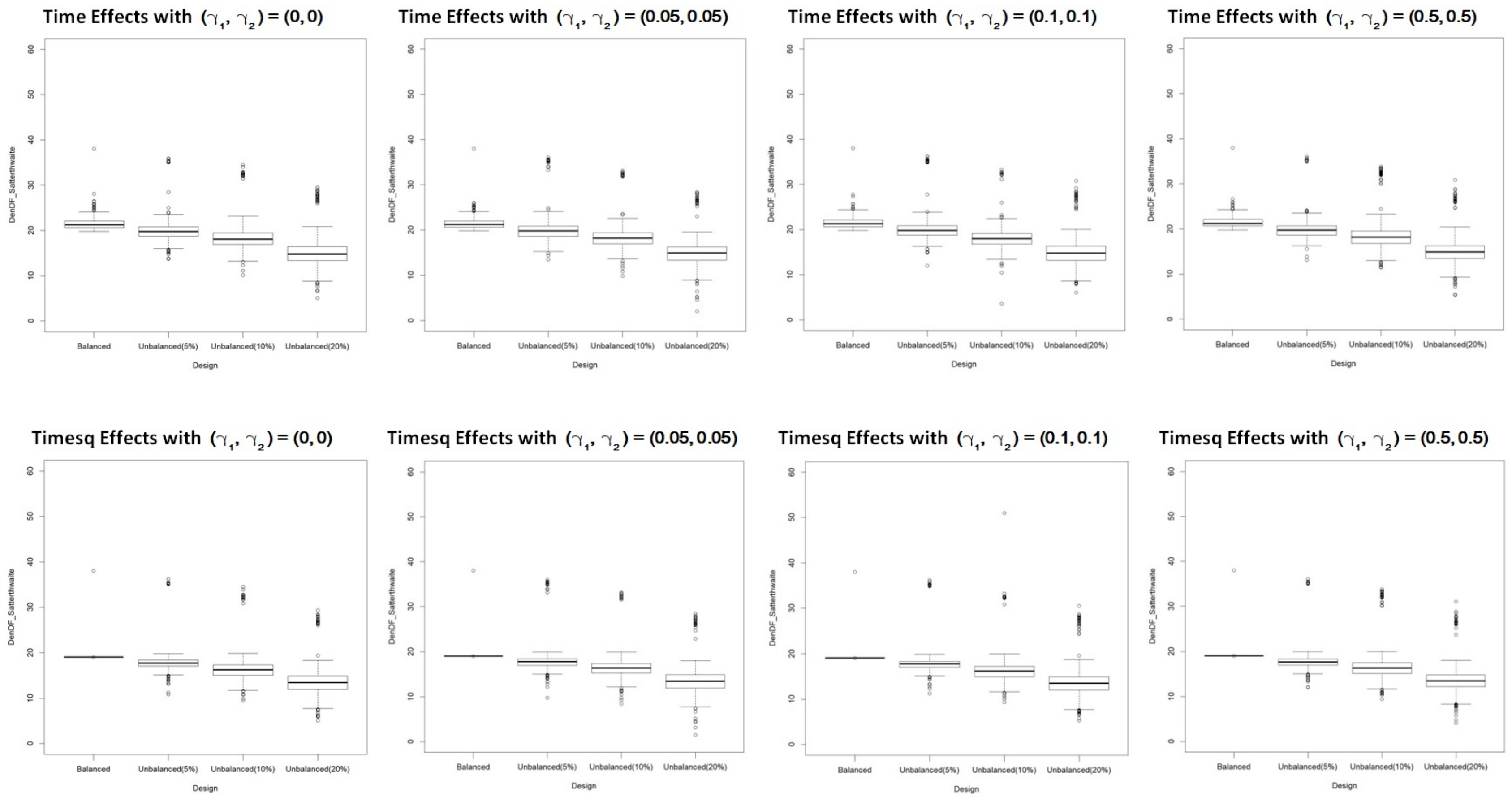


Figure 20

Satterthwaite DDF Boxplots with 20×6 Sample Structure

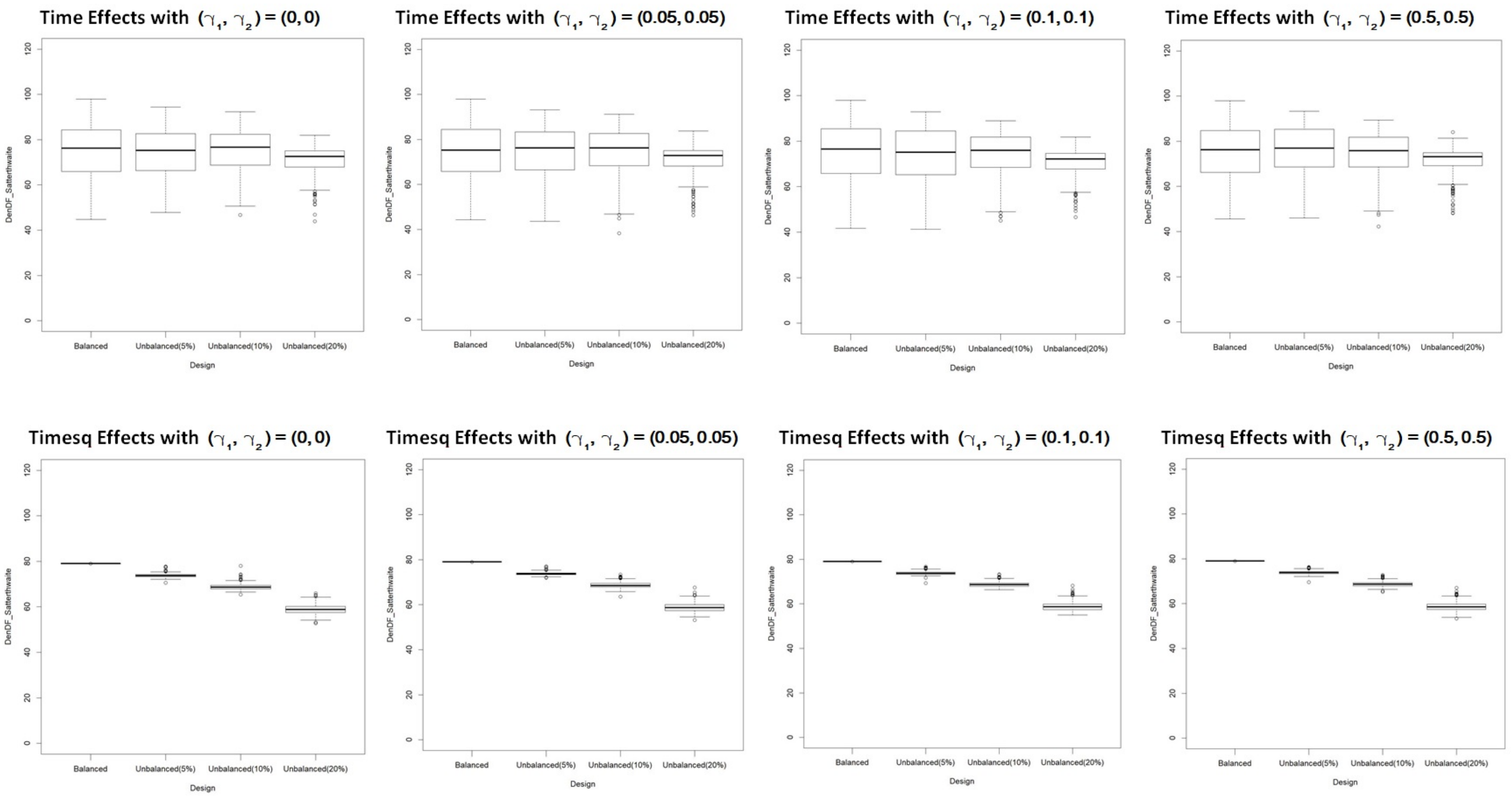
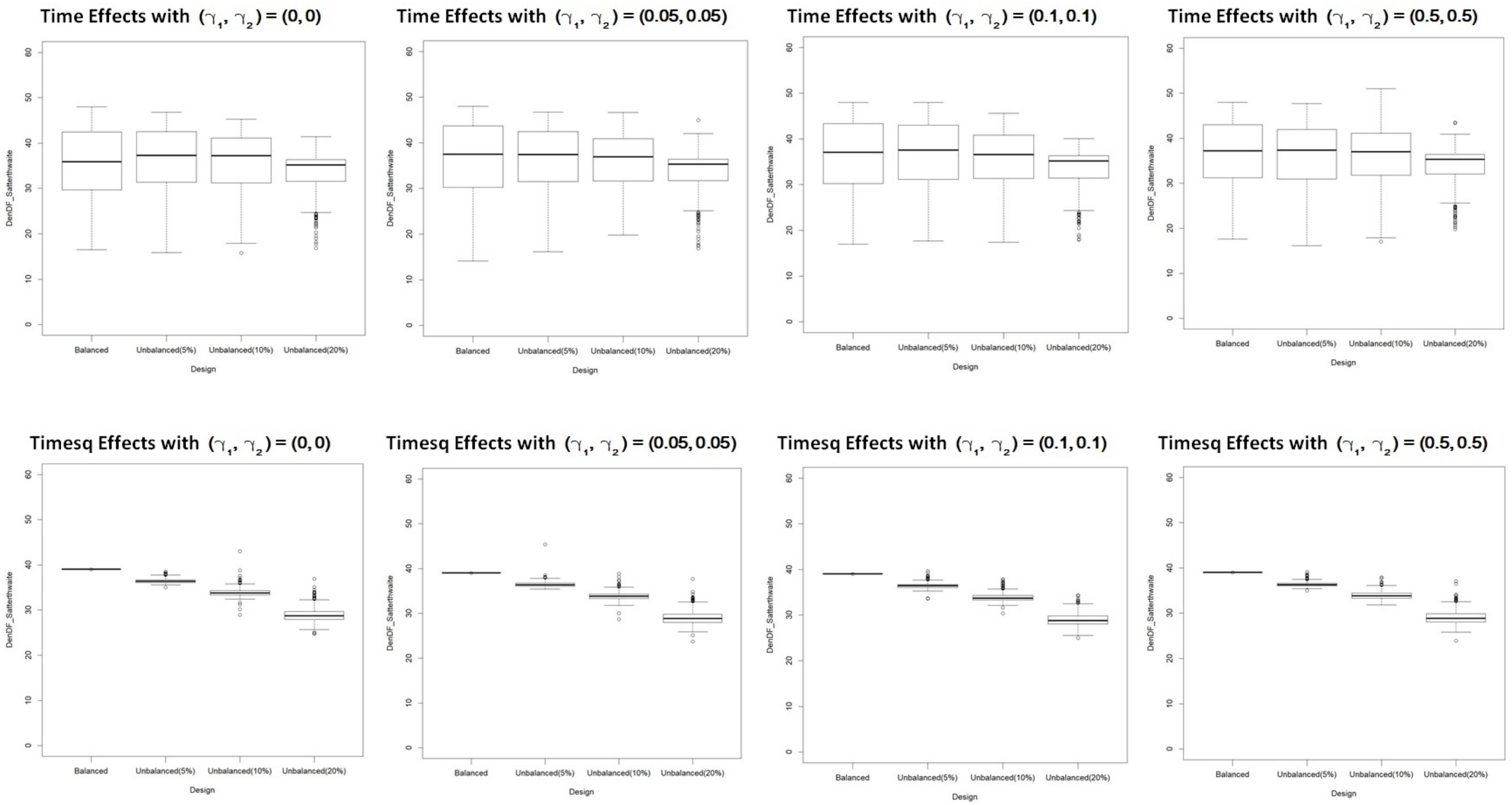


Figure 21

Satterthwaite DDF Boxplots with 10×6 Sample Structure



3.2.4.3 P-Value Histograms. We first obtain the mean percentages of p-values less than $\alpha = 0.05$ across different missing-data proportion as follows:

Table 5

Satterthwaite Mean Percentage for (γ_1, γ_2) = (0, 0)

No.	Sample Structure	Effect Type	Mean Percentage (%)
1	20×3	Time	5.7
2	20×3	Timesq	5.3
3	20×6	Time	5.2
4	20×6	Timesq	5.7
5	10×6	Time	4.9
6	10×6	Timesq	4.5

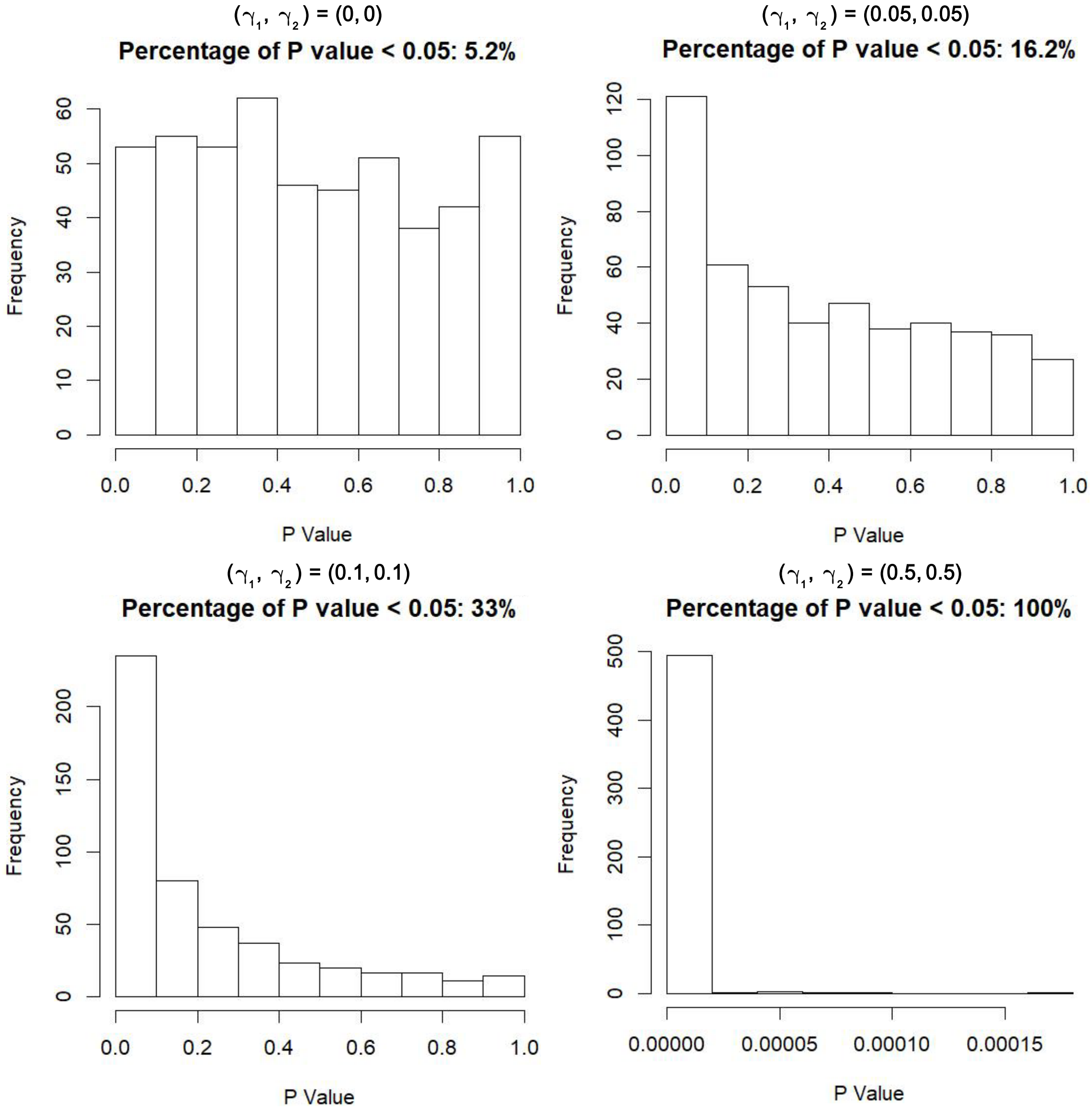
*beyond the confidence interval of (3.09%, 6.91%)

Note there is no mean percentage falling beyond the confidence interval. This result indicates that the Satterthwaite method performs well in the case of $(\gamma_1, \gamma_2) = (0, 0)$. The different settings of effects and sample structure do not significantly influence the power generated by the Satterthwaite method.

A trend of increasing power of rejecting the null hypothesis is present as illustrated in Figure 22 on the next page. This applies to all the cases of different effects, sample structure and missing-data proportion. In addition, the power is generally larger for the timesq effects than for the time effects, especially in the cases with 5% and 10% missing values.

Figure 22

Satterthwaite P-Value Histograms for 20 × 3 Unbalanced (5%)Timesq Effects



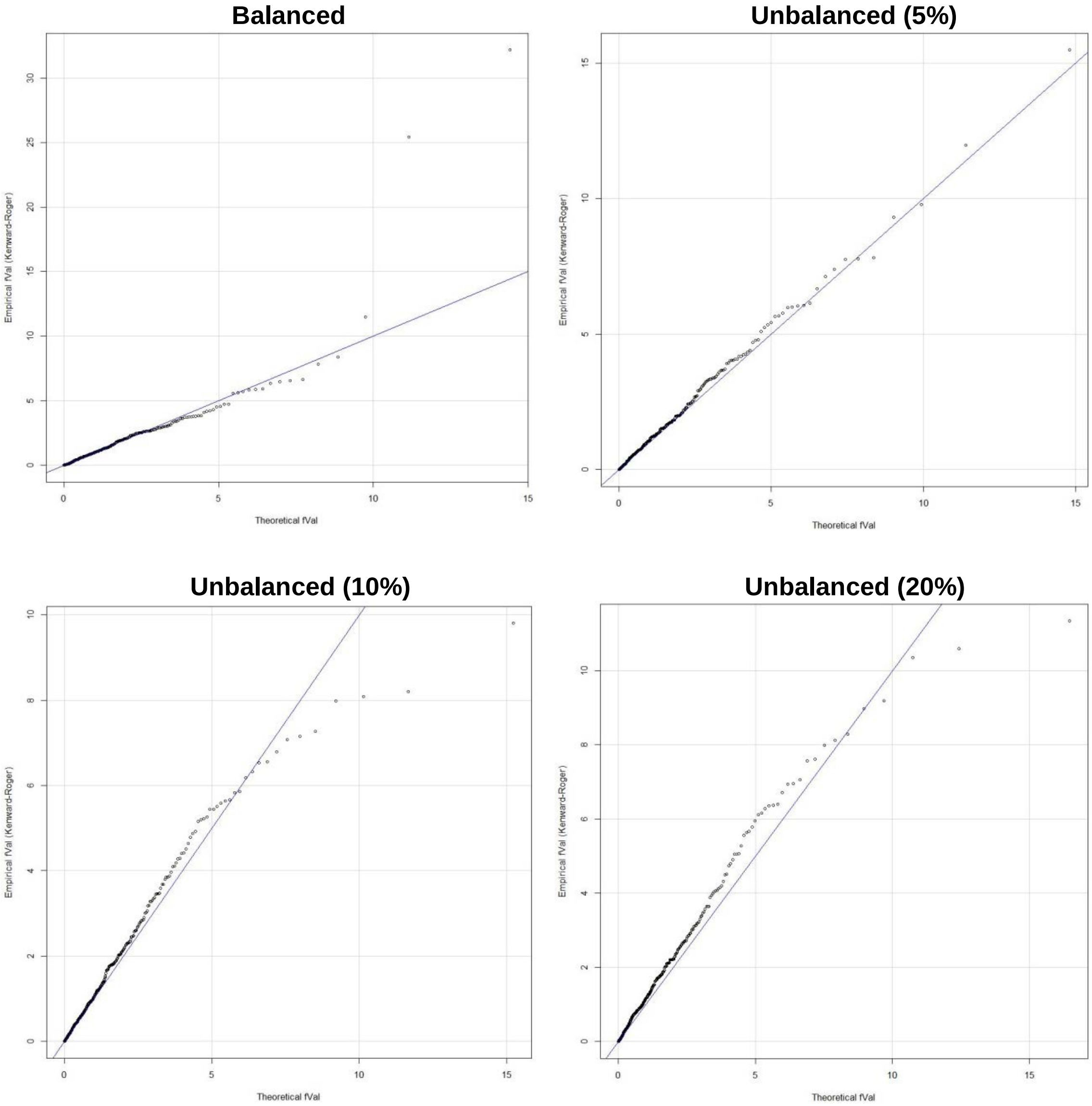
3.2.5 Kenward-Roger Method

3.2.5.1 F-Statistic Q-Q Plots. Previously there is a positive relationship between the F-Statistic deviance and missing-data proportion for the case of 20×3 sample structure by the residual, containment, B-W and Satterthwaite methods.

Nevertheless, here such relationship is not as clear (Figure 23). The deviance between the empirical and the theoretical F-Statistics for the case with 20% missing data (lower right in Figure 23) is the smallest among all the DDF approximation methods. This result is also present in the case for the 20×3 timesq effects.

Figure 23

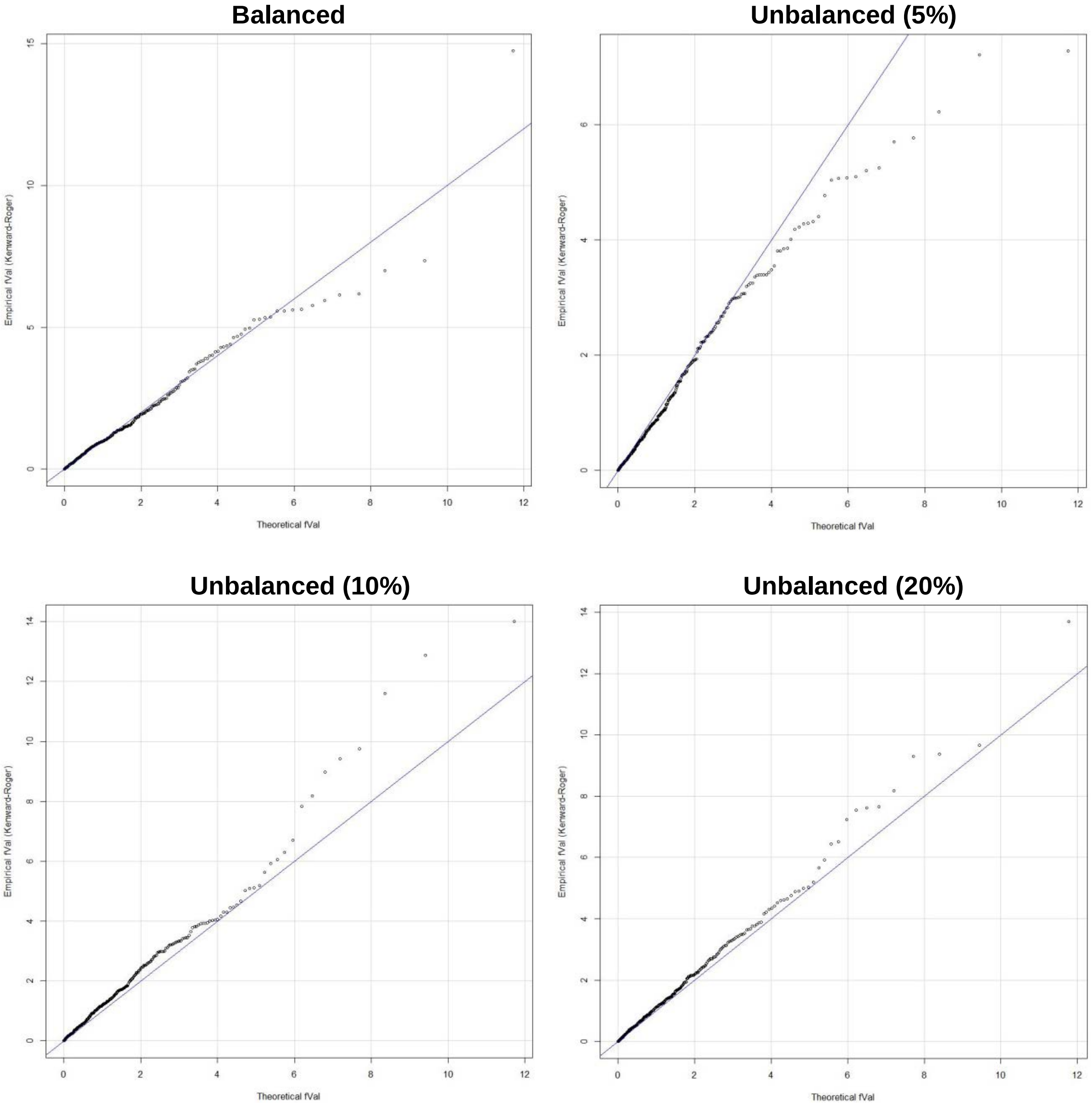
K-R F-Statistic Plots Highlighted with 20×3 Time Effects



There is no significant difference between the Q-Q plots with the 20×3 sample structure and the ones with the 20×6 sample structure. Increasing the number of time pointes nested within each factor level does not make much difference in this case like it does in the previous cases with the other DDF approximation methods.

Figure 24

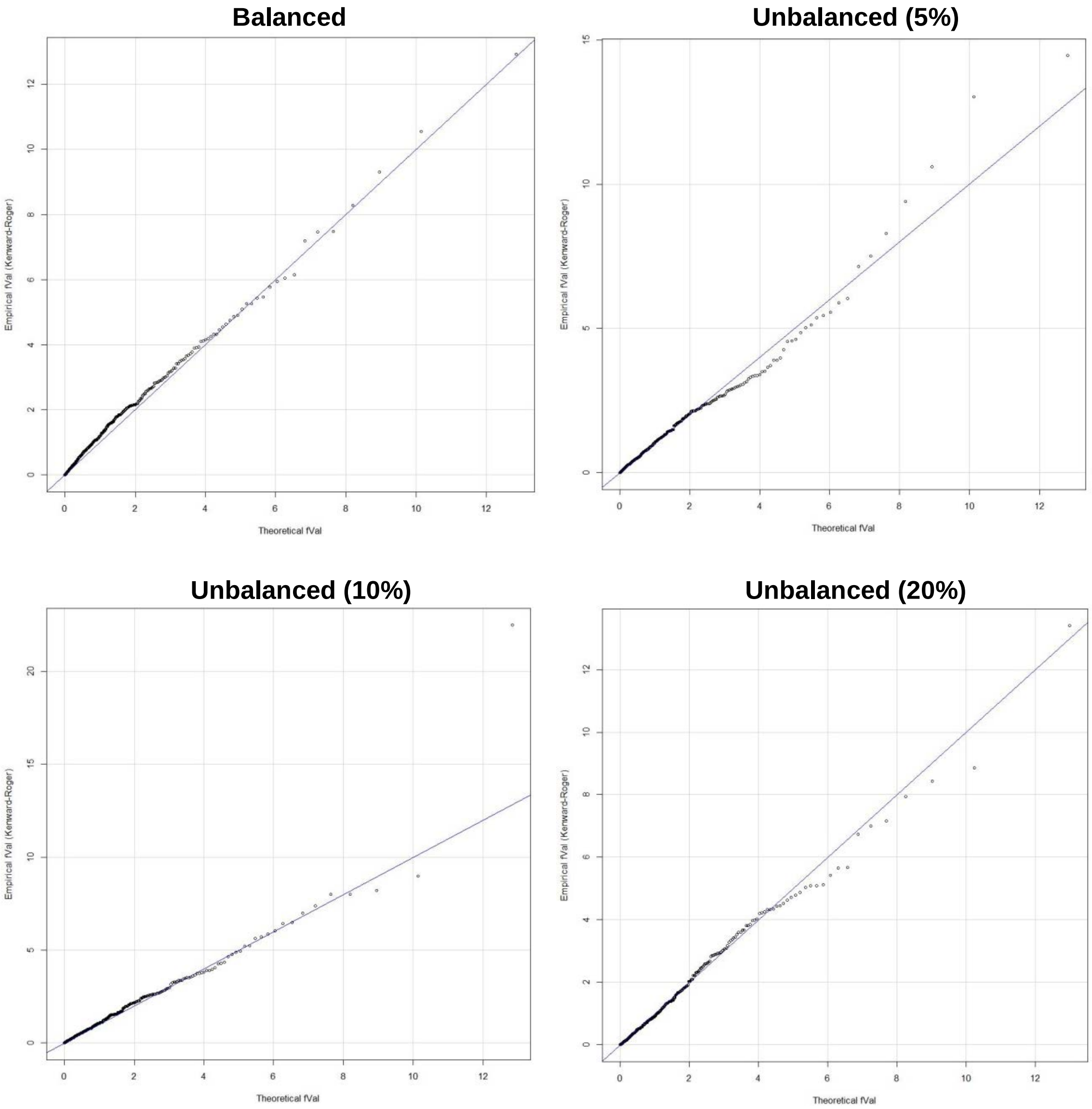
K-R F-Statistic Q-Q Plots Highlighted with 20×6 Time Effects



Nonetheless, when we further decrease the number of factor levels, an overall smaller F-Statistic deviance can be observed in Figure 25 below. The influence on the F-Statistic deviance by decreasing the number of factor levels is greater than by increasing the number of time points for the K-R method.

Figure 25

K-R F-Statistic Q-Q Plots Highlighted with 10 × 6 Time Effects



3.2.5.2 DDF Boxplots. The patterns we discover from the K-R method are similar to those from the Satterthwaite method. Contrary to the other methods that give robust DDF across different settings of slopes, sample structure and missing-data proportion, the Satterthwaite method and the K-R method are relatively sensitive to the change of designs and settings.

First of all, the patterns of DDF boxplots for the K-R method with 20×3 sample structure (Figure 26 on p.43) are quite similar to the ones for the Satterthwaite method (Figure 19 on p.34). There is a trend of decreasing mean DDF as the proportion of missing values increases. In addition, a certain amount of outliers is present and scatters in a similar fashion for the K-R method and the Satterthwaite method.

Second, in Figure 27 on p.44, the patterns, although differ between the time and timesq effects, are similar to the way they vary from the Satterthwaite method. The DDF distribution is more concentrated for the timesq effects than for the time effects. While the means for the time effects stay almost the same across different missing-data proportion, the means for the timesq effects decreases as the proportion increases.

Finally, in Figure 28 on p.45, there is relatively a longer tail among the DDF distributions for the time effects, compared to the case with 20×6 sample structure. There is almost no difference between the results from the K-R method and from the Satterthwaite method. However, note that for the cases with 20×6 and 10×6 sample structure, the DDF distribution is more scattered compared to the case with 20×3 sample structure. It may indicate that increasing the number of time points nested within each factor level increases the variability for the K-R and Satterthwaite methods.

Figure 26

K-R DDF Boxplots with 20×3 Sample Structure

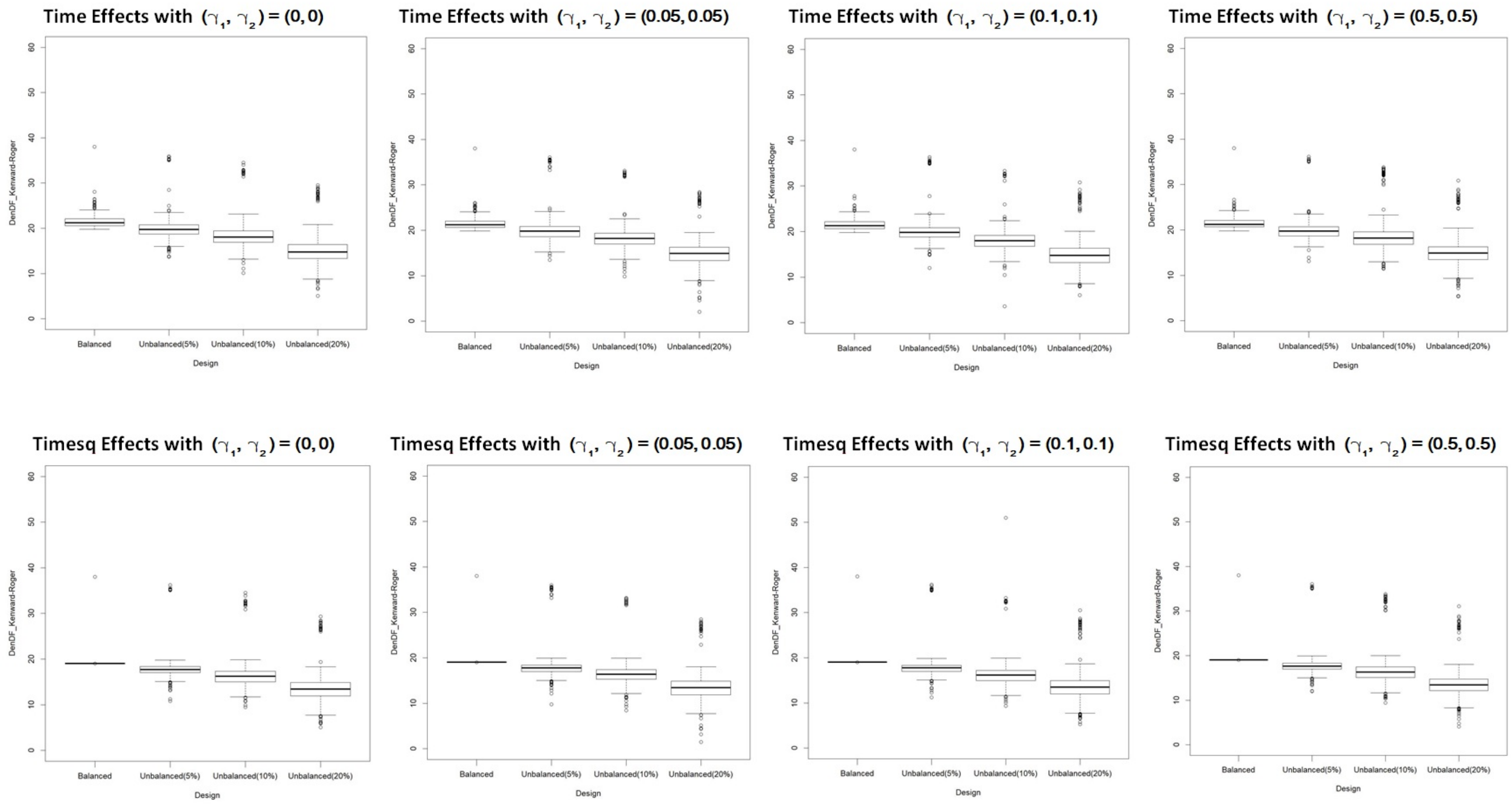
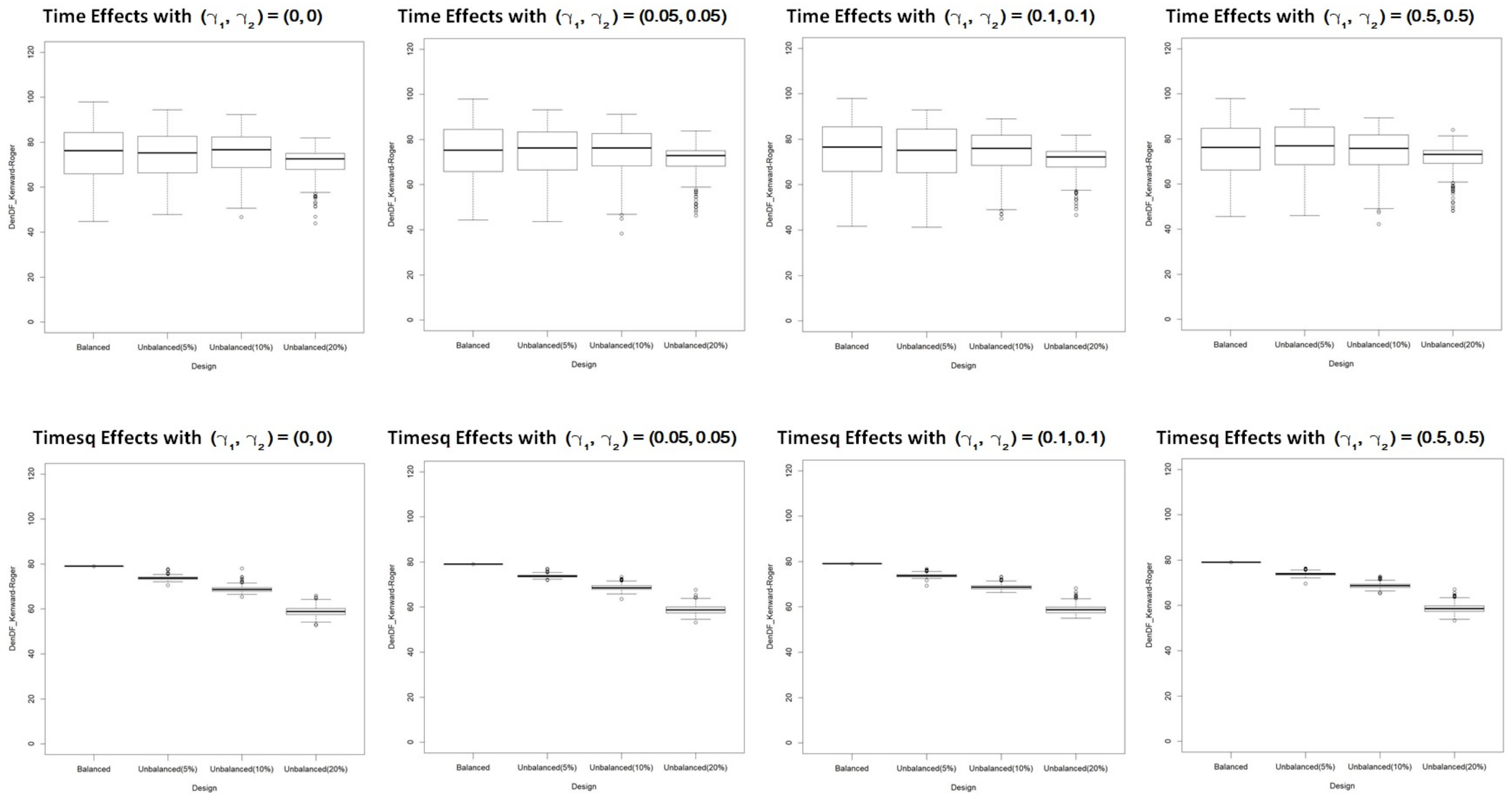


Figure 27

K-R DDF Boxplots with 20×6 Sample Structure



Time Effects with $(\gamma_1, \gamma_2) = (0, 0)$

Time Effects with $(\gamma_1, \gamma_2) = (0.05, 0.05)$

Time Effects with $(\gamma_1, \gamma_2) = (0.1, 0.1)$

Time Effects with $(\gamma_1, \gamma_2) = (0.5, 0.5)$

Timesq Effects with $(\gamma_1, \gamma_2) = (0, 0)$

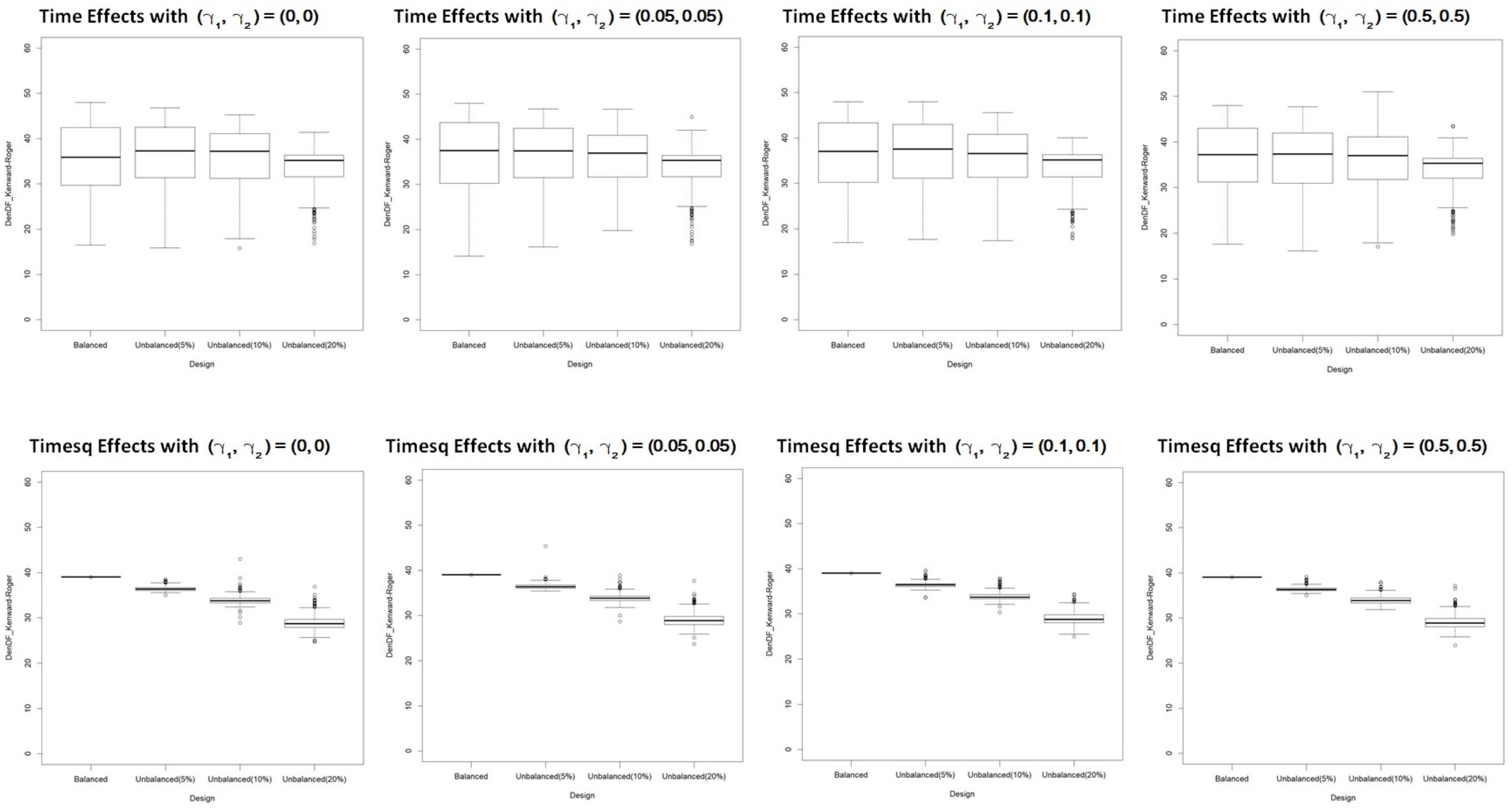
Timesq Effects with $(\gamma_1, \gamma_2) = (0.05, 0.05)$

Timesq Effects with $(\gamma_1, \gamma_2) = (0.1, 0.1)$

Timesq Effects with $(\gamma_1, \gamma_2) = (0.5, 0.5)$

Figure 28

K-R DDF Boxplots with 10 × 6 Sample Structure



3.2.5.3 P-Value Histograms.

First, we compute the mean percentages of p-values less than $\alpha = 0.05$ across different missing-data proportion as follows:

Table 6

K-R Mean Percentage for (γ_1, γ_2) = (0, 0)

No.	Sample Structure	Effect Type	Mean Percentage (%)
1	20×3	Time	5.4
2	20×3	Timesq	5.3
3	20×6	Time	5.2
4	20×6	Timesq	5.7
5	10×6	Time	4.7
6	10×6	Timesq	4.5

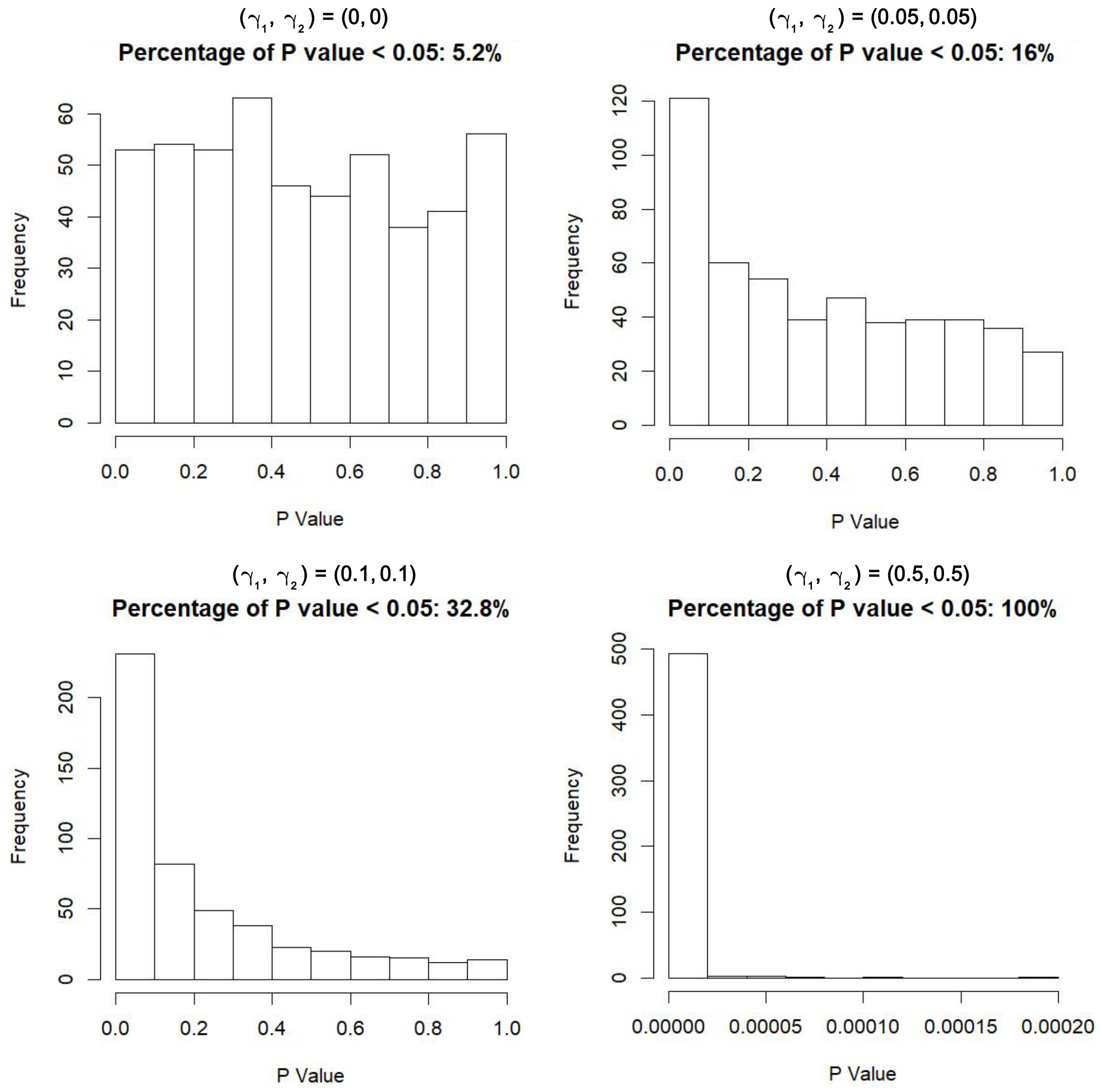
*beyond the confidence interval of (3.09%, 6.91%)

Note there is no mean percentage falling beyond the confidence interval. This result shows that the K-R method performs well in the case with $(\gamma_1, \gamma_2) = (0, 0)$ regardless of the different settings of effects and sample structure.

A universal trend of increasing power through the increase in the slopes for the time and timesq effects is present, as exemplified in Figure 29 on the next page. Additionally, similar to the previous results from the other methods, the power for the timesq effects tends to be greater than the one for the time effects, especially in the cases with 5% and 10% missing values.

Figure 29

K-R P-Value Histograms for 20×3 *Unbalanced (5%)Timesq Effects*



4 Conclusion

4.1 Discussion of Simulation Results

In this study, we aim to evaluate the performance of residual, containment, between-within, Satterthwaite, and Kenward-Roger methods of denominator degrees of freedom approximations in small-sample situations. We conduct simulations on a mixed model consisting of random intercept and random slope and look into the deviance between empirical and theoretical F-statistics, the distribution of denominator degrees of freedom, and the patterns of p-values. There are three major findings in our study:

(1) In the case where the null hypothesis is true, the Kenward-Roger method performs the best in terms of the deviance between the empirical and theoretical F-statistics. Moreover, regardless the type of methods, the deviance between the empirical and theoretical F-statistics tends to become smaller by increasing the number of observations nested within each factor level.

(2) While the other methods give denominator degrees of freedom that are robust across different settings of the sample structure, the Satterthwaite and Kenward-Roger methods are relatively sensitive to the change of designs. In particular, the variability in the computation and outcome goes up when the number of observations nested within each factor level increases.

(3) The residual method performs the worst among the five methods in terms of p-values. In addition, the power for the timesq effects tends to be greater than the one for the time effects in our simulation model.

4.2 Area for Future Exploration

One interesting thing to note is that in the process of developing our simulation, we tried out the Satterthwaite and Kenward-Roger methods (the only two approximation methods available) in R and discovered a difference between the denominator degrees of freedom produced by R and by SAS. With the SAS command type=UN, the covariance structure in R and SAS should be equivalent and the results should be quite similar. However, out of 100 data in total, there are averagely 2 cases of exceptions for every setting-variable combination. In the exceptional cases, there is a significant difference between the denominator degrees of freedom generated by the lmer function in R and the Proc Mixed procedure in SAS. The difference ranges from 1 to 29, while the average difference for the majority of data is around 0.05. This type of extreme differences tends to occur when R produces particular large or small values.

Further looking into the exceptional cases with extreme values, we observed that in R it shows warning messages about failing to converge, whereas in SAS the convergence criteria is always met. Additionally, in the exceptional case in R where the same warning messages popping up for both of the Satterthwaite and Kenward-Roger methods, Satterthwaite method tends to produce particular large or small values that are quite different from the values produced by SAS, while Kenward-Roger method may produce values that are very similar to SAS. So we conclude that the effect of the convergence problem in R varies between Satterthwaite and Kenward-Roger methods.

We then tried to change the optimizers in R (optimizer=Nelder_Mead, method=nlminb, method=L-BFGS-B), which is one of the common solutions for convergence problems. Yet all of the optimizers fail to converge and produce even more extreme values. Since the difference between R and SAS is not our focus in this study, we did not go further. Future research may be developed based on this finding and provide more insights on the application of software packages for linear mixed models and the possible issues that should be addressed for their algorism.

References

Adriaenssens, B. (2015). *Simulating a random intercept, random slope model*. Rstudio-Pubs-Static.S3.Amazonaws.Com. Retrieved June 19, 2019, from http://rstudio-pubs-static.s3.amazonaws.com/133409_8567e08df08d42a8977853e4f94ca005.html.

Diggle, P. (2002). *Analysis of longitudinal data*. Oxford University Press.

Fai, A. H., & 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*(4), 363–378. https://doi.org/10.1080/00949659608811740.

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*(2), 477–486. https://doi.org/10.2307/2530872.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association*, *72*(358), 320–338. https://doi.org/10.1080/01621459.1977.10480998.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis : techniques and applications*. Routledge, Taylor & Francis Group.

Kackar, R. N., & Harville, D. A. (1984). Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models. *Journal of the American Statistical Association*, *79*(388), 853–862. https://doi.org/10.2307/2288715.

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics*, *53*(3), 983–997. https://doi.org/10.2307/2533558.

Kreft, I. G. G., & Leeuw, J. de. (1998). *Introducing multilevel modeling*. London: SAGE.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. New York: Springer-Verlag.

Li, P., & Redden, D. T. (2015). Comparing denominator degrees of freedom approximations for the generalized linear mixed model in analyzing binary outcome in small sample cluster-randomized trials. *BMC Medical Research Methodology*, *15*(1). https://doi.org/10.1186/s12874-015-0026-x.

Luke, S. G. (2016). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y.

Mcculloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, linear, and mixed models*. John Wiley & Sons.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554. https://doi.org/10.1093/biomet/58.3.545.

Rencher, A. C., & Schaalje, G. B. (2008). *Linear Models in Statistics*. John Wiley And Sons.

SAS Institute. (2017). *SAS/STAT 14.3 User's Guide The MIXED Procedure*. Cary, NC: Author.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, *6*(5), 309–316. https://doi.org/10.1007/bf02288586.

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*(4), 512–524. https://doi.org/10.1198/108571102726.

Schluchter, M. D., & Elashoff, J. T. (1990). Small-sample adjustments to tests with unbalanced repeated measures assuming several covariance structures. *Journal of Statistical Computation and Simulation*, *37*(1–2), 69–87. https://doi.org/10.1080/00949659008811295.

Searle, S. R., Casella, G., & Mcculloch, C. E. (2006). *Variance components*. John Wiley & Sons.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London/Thousand Oaks/New Delhi: SAGE.

Tietjen, G. L. (1974). Exact and Approximate Tests for Unbalanced Random Effects Designs. *Biometrics*, *30*(4), 573. https://doi.org/10.2307/2529222.

West, B. T., Welch, K. B., Gałecki, A. T., & Gillespie, B. W. (2015). *Linear mixed models : a practical guide using statistical software*. Crc Press, Taylor & Francis Group.

APPENDIX A

PSEUDOCODE FOR DATA SIMULATION

```
Install and load packages lme4, lmerTest, and car
Initialize variables se, ti, ti_sq, group, individual, and group_individual
Create a vector (design) to store the four design types
Set omitted_value <−group_individual*design


Create folders for the results produced
Create a vector (design_name) to store the names of the four designs


For a=1 to the length of design
#Create a loop to generate time
  Set k=1
  Create time as an empty vector
  For i=1 to group
    For j=1 to individual
      Place j in the kth element of time
      k=k+1


  #Create a loop to generate timesq
  Set k_sq=1
  Create timesq as an empty vector
  For i=1 to group
    For j=1 to individual
      Place j*j in the k_sq th element of timesq
      k_sq=k_sq+1

#Create a loop for 500 datasets
  Set seed (se)
  For i=1 to 500
    Call the function as.formula to set the model that the data should follow
    Set the predictor variables
    Set the fixed effect parameters
    Set the random effect structure
    Simulate data using the parameters above by calling the simulate function
    Set up missing values by calling sample function
    Sort the result by ID and format
    Export the data as txt
```

APPENDIX B

PSEUDOCODE FOR MIXED MODELING

```
Create datasets to store results
Set up macro variable RunModel
Initialize dataset Allresult
Initialize rep
Do i=1 to 500
Read in the output generated by R
/*Process model*/
/*Repeat the following procedure for the five models within the do-loop*/
Use proc mixed procedure to process mixed modeling by residual method
Use ods output to store the output to the dataset result_R
/*Subset the results for time and timesq into two datasets*/
Create two datasets subresult_R_time and subresult_Rtimesq to store the subset data
If (effect ==“time”)
  Store the output to subresult_R_time
Else
  Store the output to subresult_R_timesq
/*Update the data*/
Create result_R_time to store the output of time produced by residual method
Save the subset output in subresult_R_time into result_R_time
Create result_R_timesq to store the output of time produced by residual method
Save the subset output in subresult_R_timesq into result_R_timesq
Drop the temporary datasets subresult_R_time and subresult_R_timesq
/*End of the do-loop*/
```

```
/*Export datasets to csv files*/

Use proc export procedure to export output of result_R_time

Use proc export procedure to export output of result_R_timesq

Use proc export procedure to export output of result_C_time

Use proc export procedure to export output of result_C_timesq

Use proc export procedure to export output of result_B_time

Use proc export procedure to export output of result_B_timesq

Use proc export procedure to export output of result_S_time

Use proc export procedure to export output of result_S_timesq

Use proc export procedure to export output of result_K_time

Use proc export procedure to export output of result_K_timesq
```

APPENDIX C

PSEUDOCODE FOR PLOT GENERATION

```
Set simulation variables se, ti, ti_sq, group, individual, and group_individual
Create a vector (design) to store the four design types
Set omitted_value <—group_individual*design

Create a vector (design_name) to store the names of the four designs
Create folders to store plots
Create a data frame to store DDF for boxplots
Create data frames to store p-values for histograms

For a=1 to the length of design
  Read in the outputs generated by SAS
  Set up Q-Q plot if (ti==0 and ti_sq==0)
    #Set up Q-Q plots for time
    Calculate the mean of DDF and assign it to avg
    Assign the generated F-statistics to EmpiricalF
    Use qqPlot method to plot Q-Q plots with parameters avg and EmpiricalF
    dev.off()
    #Set up Q-Q plots for timesq
Calculate the mean of DDF and assign it to avg
    Assign the generated F-statistics to EmpiricalF
    Use qqPlot method to plot Q-Q plots with parameters avg and EmpiricalF
    dev.off()
  #Extract data for boxplots
  Create data_temp to temporarily store the extracted read-in data
  Use cbind to extract the read-in data and save it into data_temp
  Use rbind to update the data in data_box
  #Extract data for histograms
  Use cbind function to extract and combine data by the five methods

#Plot and export boxplots for each method
Use jpeg() function to create a jpeg file
Use boxplot function to plot the data
dev.off()

#Plot and export histograms for each method
Use jpeg() function to create a jpeg file
Use par() function to partition the window into 2 by 2
Use hist() to plot for the histograms of balanced design
Use hist() to plot for the histograms of unbalanced (5%) design
Use hist() to plot for the histograms of unbalanced (10%) design
Use hist() to plot for the histograms of unbalanced (20%) design
dev.off()
```