Better Power Methods for the Univariate Approach to Repeated Measures

Matthew J. Gribbin

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics, School of Public Health.

Chapel Hill

2007

Approved by:

Chairman: Keith E. Muller

Co-Chairman: Lloyd J. Edwards

Reader: Paul W. Stewart

Reader: Hongbin Gu

Reader: Jiu-Chiuan Chen

Abstract

MATTHEW J. GRIBBIN: Better Power Methods for the Univariate Approach to Repeated Measures (Under the direction of Dr. Keith E. Muller)

New methods that improve upon current techniques related to power for UNIREP tests are introduced. The research is motivated by imaging applications, which often generate the type of data that can be handled with UNIREP techniques. The UNIREP Huynh-Feldt test is based on the Huynh-Feldt sphericity estimator. Claiming their estimator was a ratio of unbiased estimators, Huynh and Feldt developed it as an alternative to the sometimes biased Geisser-Greenhouse estimator. The Huynh-Feldt estimator is examined and shown to be a ratio of unbiased estimators only for the special case of rank of the design matrix, $\boldsymbol{X}$, equal to 1. This realization results in a biased Huynh-Feldt test and power calculation when rank of $\boldsymbol{X}$ is greater than 1. A proper, adjusted Huynh-Feldt estimator for any rank of $\boldsymbol{X}$ is presented and shown to better estimate the population sphericity when rank of $\boldsymbol{X}$ is greater than 1. A power approximation for the rank-adjusted Huynh-Feldt test is also presented. For practical research situations, the rank-adjusted Huynh-Feldt power approximation is shown to perform as well as and, in most cases, better than the most accurate Huynh-Feldt power approximation in use. Furthermore, the Huynh-Feldt power approximation is shown to yield artificially inflated power values at a cost of inflated test size when rank of $\boldsymbol{X}$ is greater than 1. The rank-adjusted test is shown to control test size adequately. Approximate confidence intervals for UNIREP power in the case of an estimated covariance and fixed means are introduced and shown to provide reasonably accurate coverage probabilities for all four UNIREP tests. The approximate confidence intervals

perform well in most cases considered, even for small sample sizes. The approximate confidence intervals are shown to perform better for higher power values than for lower power values, making them more useful in practical research conditions. Factors affecting UNIREP power confidence interval coverage probabilities are examined. These factors include sample size, rank of $\boldsymbol{X}$ and the degrees of freedom for both the estimating and target studies, as well as estimated sphericity multipliers. To provide tighter, more informative confidence bounds, one-sided confidence intervals are recommended.

To Sandy

Acknowledgements

More than anyone else in the world, I must thank my beautiful wife Sandy for her love, encouragement and friendship. The dream of a life spent loving her is what gives me the courage and determination to reach beyond my expectations. Her love and support has made me a better person than I could be by myself, and I thank God every day that she is in my life.

I also wish to thank our family for their support and sacrifices that have allowed me the opportunity to come this far. To have such loved ones in my life is truly an honor and a blessing. I hope to lead the kind of professional and personal life that will make them proud.

I owe much gratitude to Keith Muller, my dissertation advisor, mentor and friend. He has provided me with the skills and training I needed to complete my research, but more importantly, he has prepared me for a fulfilling career as a biostatistician. I must also thank my committee: Lloyd Edwards, Paul Stewart, Hongbin Gu and Jiu-Chiuan Chen. They have all provided me with whatever support I have asked from them, and each has given me invaluable insight that improved my research.

Finally, I am grateful to the National Institute of Environmental Health Sciences for allowing me the opportunity to serve as a trainee on the NIEHS Training Grant #5-T32-ES007018-30.

Table of Contents

Lists of Tables

Tables

List of Figures

Figures

Chapter 1
Introduction and Literature Review

1.1 Introduction

Multivariate analysis techniques are used when data is collected on subjects with more than one response value, either due to multiple outcomes or repeated measures. One strategy for analysis of such data is the univariate approach to repeated measures (UNIREP). A tremendous amount of work has gone into developing UNIREP methods over the past 75 years. However, there are still techniques that need to be examined and improved upon. Three areas related to power for UNIREP tests are the focus of the present research.

Although the new methods that will be introduced and discussed may be applied to any number of studies, the driving motivation and application has been imaging research. Imaging is being used more and more in all forms of medical research, and the cost of such procedures is constantly decreasing. Researchers and physicians alike are realizing the benefits to using these safe and non-invasive techniques.

UNIREP techniques make up a special case of the more broad area of statistical modeling called mixed models. The mixed model has several nice statistical features, such as no requirement for balanced data, the ability to explicitly model and analyze the between- and within-subject variation, and the capability of handling missing data without eliminating all values for a particular subject. However, there is still a need for better inference and power analysis techniques in mixed models. This is much less true for UNIREP. The inference techniques for UNIREP far outshine those used in mixed models,

particularly for small sample sizes, and power techniques for UNIREP have been well tested and documented.

The methods presented here focus solely on UNIREP procedures. The expectation is that this research will lay the groundwork for future researchers to ultimately extend these methods to fit with the general mixed model. This progression seems to be a natural one. UNIREP methods were generalized by Catellier and Muller (2000) to allow for missing data. However, no repeated covariates are allowed, and no power calculation is currently available.

Much of imaging research does not require the analysis qualities that are associated with mixed model procedures. Although in field studies missing data may be common, imaging research often generates the type of complete data that can be handled with UNIREP procedures. Although a subject may be missing and MRI from a research study, there is no data missing within an MRI, and, in UNIREP, there is no need for balance between subjects, only within. Imaging does not always provide complete data, however. Pre-imaging processing may lead to situations in which portions of the imaging data may be missing. Thus, the techniques discussed here may not be appropriate for some imaging research studies.

Although imaging research has been the driving motivating application of the current research, the overall applications extend much further. Experimental or controlled laboratory research, such as animal studies (e.g. mouse recombinant DNA) or some psychiatric studies, will often possess the type of complete data required for use of the techniques discussed here. Also, such studies often have small sample sizes, which makes UNIREP (and MULTIREP) techniques much more desirable than mixed models.

Prior to introducing new methods, familiarization with existing methods is necessary. In Chapter 1, many of the highlights pertaining to the development of UNIREP techniques

are reviewed, including UNIREP power, confidence intervals for univariate analyses and UNIREP sphericity estimators.

In 1976, Huynh and Feldt developed a new estimator of sphericity that improved upon the sometimes biased Geisser-Greenhouse estimator. They claimed that the Huynh-Feldt estimator was a ratio of unbiased estimators. The Huynh-Feldt UNIREP test uses this estimator when calculating degrees of freedom for its approximate F distribution. In Chapter 2, the Huynh-Feldt estimator is examined and shown to be a ratio of unbiased estimators only for the special case of rank of the design matrix, $\boldsymbol{X}$, equal to 1. This realization may result in a biased Huynh-Feldt test and power calculation when rank of $\boldsymbol{X}$ is greater than 1. A proper estimator for any rank of $\boldsymbol{X}$ is presented and evaluated for a wide range of conditions.

Improving upon existing power calculation methods, Muller *et al.* (2007) introduced approximate power calculations for all four UNIREP tests which were accurate and easy to use. Their power approximation for the Huynh-Feldt test incorporates the Huynh-Feldt sphericity estimator. In Chapter 3, the work begun in Chapter 2 is extended by incorporating the rank-adjusted approximately unbiased estimator into a power approximation, similar to the Muller *et al.* (2007) power approximation. The accuracy of the rank-adjusted power approximation is evaluated for a wide range of conditions, and the rank-adjusted test is shown to control test size as rank of $\boldsymbol{X}$ increases better than the Huynh-Feldt test.

Accurate power analysis is essential when designing a study. Accurate power analysis allows researchers the ability to focus the study hypotheses, clarify the analysis plans and enhance study design efficiency. When variance is estimated, power becomes a random variable. Providing confidence intervals that account for the uncertainty in these random power values would be useful in any study design. For instance, a lower bound for power

would allow a researcher to state that a study has power of at least "P" to detect an effect, with a specified confidence.

For estimated variance and fixed means, exact power confidence intervals for univariate analyses have been presented by Taylor and Muller (1995). In Chapter 4, the methods of Taylor and Muller (1995) are extended to provide accurate, approximate confidence intervals for UNIREP power. The methods are evaluated using simulations employing a wide range of conditions, and are shown to be accurate enough for any power analysis.

1.2 Notation

A column vector $\boldsymbol{x}$, $(n \times 1)$, is lower case bold. A matrix, $\boldsymbol{X}$, is upper case bold with transpose $\boldsymbol{X}'$, inverse $\boldsymbol{X}^{-1}$ and generalized inverse $\boldsymbol{X}^{-}$. Also, $\mathbf{1}_n$ is an $(n \times 1)$ vector of 1's and $\boldsymbol{I}_n$ is an $(n \times n)$ identity matrix. A diagonal matrix with (i, i) element x_i is written $\mathrm{Dg}(\boldsymbol{x})$. The largest eigenvalue (or characteristic root) of $\boldsymbol{X}$ is $\mathrm{ch}_{\max}(\boldsymbol{X})$ and the determinant of $\boldsymbol{X}$ is $\left|\boldsymbol{X}\right|$. The expected value, the variance and the trace of $\boldsymbol{X}$ are denoted by $\mathrm{E}(\boldsymbol{X})$, $\mathcal{V}(\boldsymbol{X})$ and $\mathrm{tr}(\boldsymbol{X})$, respectively. Throughout, $X \sim \chi^2(\nu, \omega)$ indicates that the random variable X has a noncentral chi-square distribution with ν degrees of freedom and noncentrality ω. Also, $X \sim \chi^2(\nu)$ indicates that the random variable X has a central chi-square distribution with ν degrees of freedom. Similarly, $X \sim F(\nu_1, \nu_2, \omega)$ indicates X has a noncentral F distribution with ν_1 numerator and ν_2 denominator degrees of freedom, and noncentrality ω with cumulative distribution function $F_F(\nu_1, \nu_2, \omega)$. As with the central chi-square, if $\omega = 0$, X follows a central F distribution, $X \sim F(\nu_1, \nu_2)$, with ν_1 numerator and ν_2 denominator degrees of freedom. The quantile q of a central chi-square distribution with ν degrees of freedom is indicated by $F^{-1}_{\chi^2}(q; \nu)$. Similarly, the quantile q of a central F distribution is indicated by $F^{-1}_F(q; \nu_1, \nu_2)$. Furthermore, $\boldsymbol{y} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates $\boldsymbol{y}$, $(p \times 1)$, is Gaussian with mean $\boldsymbol{\mu}$ and a fixed, unknown and positive definite or positive semidefinite covariance among response variables, $\boldsymbol{\Sigma}$, $(p \times p)$. If $\boldsymbol{Y}$, $(N \times p)$, has N

independent rows and $[\text{row}_i(\boldsymbol{Y})]' \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, then $\boldsymbol{S} = \boldsymbol{Y}'\boldsymbol{Y} \sim \mathcal{W}_p(N, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ indicates $\boldsymbol{S}$ follows a Wishart distribution with N degrees of freedom, covariance $\boldsymbol{\Sigma}$, and noncentrality $\boldsymbol{\Omega} = \text{E}(\boldsymbol{Y}')\text{E}(\boldsymbol{Y})\boldsymbol{\Sigma}^{-1}$.

The General Linear Multivariate Model (GLMM),

$$\underset{(N \times p)}{\boldsymbol{Y}} = \underset{(N \times q \times p)}{\boldsymbol{XB}} + \underset{(N \times p)}{\boldsymbol{E}} , \tag{1}$$

assumes N independent rows and $[\text{row}_i(\boldsymbol{Y})]' \sim \mathcal{N}_p\big([\text{row}_i(\boldsymbol{X})\boldsymbol{B}]', \boldsymbol{\Sigma}\big)$. Equivalently, $\boldsymbol{Y} \sim \mathcal{N}_{N,p}(\boldsymbol{XB}, \boldsymbol{I}_N, \boldsymbol{\Sigma})$ indicates that $\boldsymbol{Y}$ is distributed multivariate normal with expected value $\boldsymbol{XB}$, homogeneity of covariance across rows, independence of rows (i.e. independence of the N sampling units) and Gaussian observations for the p response values. In the model, $\boldsymbol{X}$ is the fixed, known design matrix, and $\boldsymbol{B}$ represents the fixed, unknown regression coefficients. The associated general linear hypothesis is

$$\text{H}_0 : \boldsymbol{\Theta} = \boldsymbol{CBU} = \boldsymbol{\Theta}_0 , \tag{2}$$

such that $\boldsymbol{C}$, $(a \times q)$, considers the between-subject effects (rank $= a$) while $\boldsymbol{U}$, $(p \times b)$, considers the within-subject effects (rank $= b$). Without loss of generality, assume $\boldsymbol{\Theta}_0 = \boldsymbol{0}$. The unscaled noncentrality is defined as $\boldsymbol{\Delta} = (\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)'\boldsymbol{M}^{-1}(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)$, $(b \times b)$, such that $\boldsymbol{M} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}'$, $(a \times a)$. The scaled noncentrality is defined as $\boldsymbol{\Omega} = \boldsymbol{\Delta}\boldsymbol{\Sigma}_*^{-1}$, $(b \times b)$, such that, $\boldsymbol{\Sigma}_* = \boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{\Upsilon}\text{Dg}(\boldsymbol{\lambda})\boldsymbol{\Upsilon}'$, $(b \times b)$, is the covariance matrix among the transformed (hypothesis) variables, with $\boldsymbol{\Upsilon}$, $(b \times b)$, the eigenvectors of $\boldsymbol{\Sigma}_*$, such that $\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}' = \boldsymbol{\Upsilon}'\boldsymbol{\Upsilon} = \boldsymbol{I}_b$ and $\boldsymbol{\lambda}$ the vector of eigenvalues, λ_i, for $\boldsymbol{\Sigma}_*$. Corresponding estimates are $\widehat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ (when applicable, else $\widetilde{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$), $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{Y}'[\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}']\boldsymbol{Y}/\nu_e$, such that $\nu_e = N - r$, the error degrees of freedom, with $r = \text{rank}(\boldsymbol{X})$, $\widehat{\boldsymbol{\Theta}} = \boldsymbol{C}\widehat{\boldsymbol{B}}\boldsymbol{U}$, $\widehat{\boldsymbol{\Delta}} = (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \sim \mathcal{W}_b(a, \boldsymbol{\Sigma}_*, \boldsymbol{\Omega})$ and $\widehat{\boldsymbol{\Sigma}}_* = \boldsymbol{U}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{U}$, with $\nu_e\widehat{\boldsymbol{\Sigma}}_* \sim \mathcal{W}_b(\nu_e, \boldsymbol{\Sigma}_*)$. The sum of squares hypothesis matrix is $\boldsymbol{S}_H = \widehat{\boldsymbol{\Delta}}$ and the sum of squares error matrix is $\boldsymbol{S}_E = \nu_e\widehat{\boldsymbol{\Sigma}}_*$, which are independent of one another. Only testable hypotheses are considered. Testable hypotheses require full rank $\boldsymbol{C}$ and $\boldsymbol{U}$, and

$\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{C}$. This notation generally follows that presented in Muller and Stewart (2006).

1.3 Literature Review Introduction

Many strategies for analyzing multivariate data are special cases of the general linear multivariate model. The multivariate analysis of variance (MANOVA), multivariate approach to repeated measures (MULTIREP), univariate approach to repeated measures (UNIREP) and the mixed model are but a few. MULTIREP analyses model means and allow for an unstructured covariance matrix. Initially, UNIREP analyses required the assumption of compound symmetry of covariance. Approximate UNIREP tests such as the Geisser-Greenhouse and Huynh-Feldt allow the methods to be applied to all covariance structures. Mixed models allow researchers to specify the type of covariance structure desired. In this respect, they are very convenient to use. However, mixed models techniques often fall short of the MULTIREP and UNIREP techniques in terms of inference and power techniques. This research focuses on UNIREP analyses and related power calculations.

There are four UNIREP tests. They are: 1) the uncorrected (Box, 1954a, b), 2) the Huynh-Feldt (1976), 3) the Geisser-Greenhouse (1958, 1959) and 4) the Box conservative (Geisser and Greenhouse, 1958). All are computed in terms of the estimated hypothesis sums of squares, $\widehat{\boldsymbol{\Delta}}$, and the estimated variance, $\widehat{\boldsymbol{\Sigma}}_* = \boldsymbol{U}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{U}$, and all use the same test statistic, $T_u = [\text{tr}(\widehat{\boldsymbol{\Delta}})/a]/[\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)]$.

In building covariance models, the two special patterns of sphericity and compound symmetry play important and related, but distinct roles. Sphericity requires all variances equal and all covariances zero. More generally, compound symmetry requires all variances equal and all covariances equal, but not necessarily zero.

In UNIREP theory, compound symmetry of $\boldsymbol{\Sigma}$, the $(p \times p)$ covariance of $\text{row}_i(\boldsymbol{E})'$, provides a sufficient condition to allow, but not guarantee, the UNIREP test statistic to

exactly follow an $F(ab, b\nu_e)$ distribution under the null. Guaranteeing the result also requires one of two side conditions for the $(p \times b)$ contrast matrix $\boldsymbol{U}$: either 1) $b = 1$ or 2) $1 < b \leq p$ and $\boldsymbol{U}'\mathbf{1}_p = \mathbf{0}$ and $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_b$. Alternately, the weaker restriction of sphericity of $\boldsymbol{\Sigma}_* = \boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U}$, the $(b \times b)$ covariance of $\text{row}_i(\boldsymbol{E}\boldsymbol{U})'$, defines the necessary and sufficient condition, $\boldsymbol{\Sigma}_* = \sigma_*^2\boldsymbol{I}_b$. Huynh and Feldt (1970) explicated the second statement, which allows for a wider range of conditions. The special case of compound symmetric $\boldsymbol{\Sigma}$ reduces the necessary and sufficient condition to requiring $b = 1$ or $1 < b < p$ and $\boldsymbol{U}'\mathbf{1}_p = \mathbf{0}$ and $\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_b$. A compound symmetric covariance may be written $\boldsymbol{\Sigma} = \sigma^2\big[\boldsymbol{I}_p(1-\rho) + \mathbf{1}_p\mathbf{1}_p'\rho\big]$. Choosing $\boldsymbol{U}_0 = \mathbf{1}_p/\sqrt{p}$ gives $\boldsymbol{U}_0'\boldsymbol{\Sigma}\boldsymbol{U}_0 = \sigma_{*0}^2$, and $\boldsymbol{U}_t$, $(p \times b)$, such that $b = p - 1$, $\boldsymbol{U}_t'\mathbf{1}_p = \mathbf{0}$ and $\boldsymbol{U}_t'\boldsymbol{U}_t = \boldsymbol{I}_b$, gives $\boldsymbol{U}_t'\boldsymbol{\Sigma}\boldsymbol{U}_t = \sigma_{*t}^2\boldsymbol{I}_b$. Here, $\sigma_{*0}^2 = \sigma^2[1 + (p-1)\rho]$ and $\sigma_{*t}^2 = \sigma^2(1-\rho)$ are the two distinct eigenvalues of compound symmetric $\boldsymbol{\Sigma}$. The second eigenvalue has multiplicity of $(p-1)$. As for any covariance matrix, the set of p eigenvalues are the variances of the underlying principal components which have zero covariances and correlations.

If sphericity holds, then the uncorrected (UN) test is uniformly most powerful among similarly invariant tests, of exact size alpha. If sphericity is not met, the corrected UNIREP test statistics, the Huynh-Feldt (HF), the Geisser-Greenhouse (GG) and the Box conservative (Box), approximately follow a central F distribution under the null, $T_u \sim F(ab\epsilon, b\nu_e\epsilon)$. The four tests differ only by their degrees of freedom by way of different estimates of the measure of sphericity, $\epsilon = [\text{tr}^2(\boldsymbol{\Sigma}_*)]/[b\text{tr}(\boldsymbol{\Sigma}_*^2)]$. This measure of sphericity is bounded between $1/b$ and 1. The test sphericity estimator multipliers are always ordered

$$\begin{array}{ccccccc} \text{Box} & & \text{GG} & & \text{HF} & & \text{UN}, \\ 1/b & \leq & \widehat{\epsilon} & \leq & \widetilde{\epsilon} & \leq & 1 \end{array}$$

such that the uncorrected and the Box conservative tests have constant sphericity

multipliers, while the Geisser-Greenhouse and the Huynh-Feldt tests use random multipliers.

Sections 1.4 and 1.5 contain discussions of the history of MULTIREP tests and mixed models, respectively. The material is required in order to assess the relative advantages and disadvantages of the nearest competitors and best alternatives to UNIREP analysis, which is the focus of the present work. Hence, although necessary to demonstrate the viability and appeal of the UNIREP approach in comparison to competitors for many important applications, some readers may wish to skip the sections. Section 1.6 contains a discussion of the history of UNIREP tests in a similar fashion to that presented in sections 1.4 and 1.5 for MULTIREP and mixed models. The papers relating to the development and use of the Huynh-Feldt estimator are of particular relevance to the current research presented in Chapter 2: A More General Version of the Huynh-Feldt Sphericity Estimator. Section 1.7 contains a review of previous work related to power for UNIREP tests. The papers relating to the development and use of the Huynh-Feldt power approximation are utilized in Chapter 3: Approximate Power for a More General Version of the Huynh-Feldt Test. Section 1.8 contains a review of previous work related to confidence intervals for power. Nearly all of the papers in both sections 1.7 and 1.8 provide the groundwork for the current research presented in Chapter 4: Power Confidence Intervals for UNIREP Tests.

1.4 A History of MULTIREP Tests

This section contains a review of work performed towards the development of MULTIREP tests. The review is mostly a historical one, with the intention of highlighting similarities and differences with UNIREP and mixed model methods.

Smith *et al.* (1962) described the basics of MULTIREP analyses. They claimed that, at the time their paper was written, the usual approach to multivariate problems was to simply ignore, and thus fail to exploit, the correlations that may exist between responses. This approach was primarily due to the fact that computers capable of handling such calculations

were not readily available. Smith *et al.* (1962) discussed model setup, hypothesis testing (including matrices $\boldsymbol{C}$ and $\boldsymbol{U}$), hypothesis and error sums of squares, three MULTIREP test statistics and their rejection criteria. The tests they discussed were Roy's Largest Root, Wilks' Likelihood Ratio and the Hotelling-Lawley trace. Schatzoff (1966) discussed similar topics for a fourth MULTIREP test, the Pillai-Bartlett trace.

$$\begin{aligned}
&\text{1) Roy's Largest Root:} && \text{RLR} &&= \text{ch}_{\max}[(\boldsymbol{S}_H)/(\boldsymbol{S}_H+\boldsymbol{S}_E)] && (3)\\
&\text{2) Wilks' Likelihood Ratio:} && \text{WLK} &&= \big|\boldsymbol{S}_E\big|/\big|\boldsymbol{S}_H+\boldsymbol{S}_E\big| && (4)\\
&\text{3) Hotelling-Lawley:} && \text{HLT} &&= \text{tr}\big(\boldsymbol{S}_H\boldsymbol{S}_E^{-1}\big) && (5)\\
&\text{4) Pillai-Bartlett:} && \text{PBT} &&= \text{tr}\big[\boldsymbol{S}_H(\boldsymbol{S}_H+\boldsymbol{S}_E)^{-1}\big] && (6)
\end{aligned}$$

Before the mid 1960's, power for the MULTIREP tests seemed incalculable, due to the noncentral distributions of the test criterion not being expressed in a numerically feasible form. By first deriving the noncentral Wishart distribution density function, Constantine (1963) was able to derive the distributions for the MULTIREP test statistics. For the nonnull case, he suggested that $\boldsymbol{S}_H \sim \mathcal{W}_b(a, \boldsymbol{\Sigma}_*, \boldsymbol{\Omega})$ and $\boldsymbol{S}_E \sim \mathcal{W}_b(\nu_e, \boldsymbol{\Sigma}_*)$. Posten and Bargmann (1964) developed an asymptotic expansion of the distribution of $m \cdot \log(\text{WLK})$ in the form of an infinite series of weighted chi-square distributions. This allowed for the ability to approximate power.

Sugiura and Fujikoshi (1969) expanded upon the work of Constantine (1963) by developing an asymptotically correct χ^2 mixture approximation for both WLK and HLT for the nonnull case up to the order m^{-2}. Lee (1971) derived the asymptotic formula for the PBT statistic. Using the asymptotic formulae, he compared the powers of the three tests numerically, and showed that exact powers for all three tests could be calculated in the case of $p = 2$.

For moderately large sample sizes and small to moderate deviations from the hypothesis, Lee (1971) showed that no one test is superior in terms of power without specifying the alternatives. If the λ_i's are very unequal, HLT $>$ WLK $>$ PBT in terms of power. However, if the λ_i's are nearly equal, the reverse is true. The PBT power varied the

most with different alternatives, followed by the WLK, then the HLT, which varied the least.

John (1971) expanded upon the work of Posten and Bargmann (1964) and Lee (1971) by evaluating powers for the various MULTIREP tests. He also concluded that there really was no "best" test.

Olson (1974) examined all four MULTIREP tests by means of power comparisons for various examples. Based on his results, he suggested that RLR should be avoided in order to protect against nonnormality and heterogeneity of covariance. He recommended using PBT because he found PBT to be the most robust of the MULTIREP tests, while also possessing adequate power.

Olson (1974) presented several special cases. He showed that when $p = 1$, or when $s = \min(a, b) = 1$, all the MULTIREP tests are equivalent and the usual F test is the uniformly most powerful test, invariant with respect to linear transformations. In general, when $\min(p, s) > 1$, no invariant test is uniformly most powerful. Finally, he showed that when only one non-zero root exists, the power for the tests are ordered RLR $\geq$ HLT $\geq$ WLK $\geq$ PBT, with the order reversed in the diffuse situation. Using examples, he observed that the power differences between PBT, WLK and HLT were not large, in the latter case. Olson (1974) recommended always using the second ordering because, for the first ordering to win out, there must be an extremely concentrated structure. Olson (1974) noted that this type of structure is not often seen in practice.

In a paper published in 1976, Olson furthered his work by evaluating the MULTIREP tests with respect to both power and robustness. He cited previous papers in which departures from normality in the direction of positive kurtosis had relatively mild effects on type I error rates for the four MULTIREP tests. Also, these effects tended to be conservative. In cases of nonnormality, the tests were ordered PBT $\geq$ WLK $\geq$ HLT $\geq$ RLR in terms of type I error rates, with PBT remaining the closest to the nominal α level

and RLR falling furthest below. Olson (1976) suggested that departures from homogeneity of covariance produced more dramatic effects. The RLR test was most prone to an excessively high type I error rate. Although HLT and PBT did not perform well either, PBT generally resulted in the smallest increases of type I error rates. Overall, Olson (1976) reconfirmed his choice of the PBT statistic, and did so again in his 1979 paper. Stevens (1980) acknowledged and supported Olson's thoughts on the developing importance of good power analysis in multivariate analysis, but questioned his choice of test statistic.

Nagarsenker and Suniaga (1983) claimed to provide formulas that allowed accurate calculations of the WLK test statistic. However, the formulas do not work as given.

Muller and Peterson (1984) reviewed approximations previously available for noncentral distribution functions of multivariate test statistics. They acknowledged that, in practice for the null case, approximations based on an F distribution had been used with great success. Muller and Peterson (1984) extended the work in this area by providing new and numerically feasible approximations for all MULTIREP tests except RLR, based on single noncentral F random variables. They showed that the power estimates obtained from such approximations appeared to provide nearly two digits of accuracy.

Barton and Cramer (1989) approached the problem of data missing at random using the WLK test by means of the EM algorithm. They found that this method would not reduce power very much when there were a large number of observations and small amounts of missing data, but could be expensive in terms of inflated test sizes.

Muller *et al.* (1992) evaluated methods for power calculation for several examples using both MULTIREP and UNIREP tests. The authors discussed how Muller and Peterson (1984) and Muller and Barton (1989) had made power calculations for multivariate analyses convenient for MULTIREP and UNIREP, respectively. Further details regarding the results with respect to UNIREP tests are discussed in section 1.7. Muller *et al.* (1992) highlighted several special cases. For example, when $b = 1$, all MULTIREP and UNIREP tests are

equivalent providing a uniformly most powerful test, and when $b > 1$ and $a = 1$, all MULTIREP statistics transform exactly to a noncentral F. They noted that no one test is uniformly most powerful for $s = \min(a, b) > 1$ and unstructured covariance. Without a uniformly most powerful test, the choice of test depends on the alternative and the degree to which sphericity is not met.

Much like Barton and Cramer (1989), Catellier and Muller (2000) considered the problem of missing data using MULTIREP techniques. While previous papers had worked on methods for estimation for repeated measures with missing data, Catellier and Muller (2000) approached the problem with a focus towards inference. They described analogues of PBT and HLT which allowed for missing data. The authors noted that while asymptotic methods work well for large samples, seriously inflated type I error rates may exist in small samples. For all tests, accuracy decreased with more repeated measures, fewer subjects, more missing data and higher correlation within subjects. However, with no missing data the MULTIREP tests controlled the type I error rate at or below the nominal rate, even for small samples.

1.5 A History of Mixed Models

This section contains a review of work performed towards the development of mixed models. The review is mostly a historical one with the intention of highlighting similarities and differences with MULTIREP and mixed model methods.

Before the early 1980's, mixed model analyses were not commonly used in practice. However, this trend was not because the methods had not yet been developed. Rather, the reason mixed model analyses were not often used was because computers required to handle the iterative computations needed for such methods were not yet readily available. As a result of this historical fact, much of the work prior to the 1980's dealt with attempting to modify the standard multivariate techniques of the time to accommodate the various data problems commonly found in practice. These problems included unbalanced data, missing

data and/or mistimed data, for example. For instance, Rao (1972) presented a comprehensive review of past work on MULTIREP tests. In terms of estimation and hypothesis testing, he observed that very little work on missing or incomplete data analysis had been done, despite the commonality of such data in practice. Still, the theory did exist and had, in some form, since the 1930's, but was simply waiting for the computing capabilities to apply it.

In 1967, Hartley and Rao developed a procedure for maximum-likelihood estimation of the unknown constants and variances in mixed model analyses. Tests of hypotheses and confidence regions were also derived. Beale and Little (1975) offered various algorithms as alternatives to MULTIREP tests for multivariate analyses with missing data. These are just a few of the many statisticians that laid the groundwork for the mixed model techniques used today.

With the 1980's came the capability for many to apply the general theory of mixed models in common practice, due to more readily available computers. Laird and Ware (1982) helped to popularize the theory of mixed models. They approached mixed models by way of a two-stage process: first the individual, then the population. For $\boldsymbol{y}_i$, $(n_i \times 1)$, $\boldsymbol{X}_i$, $(n_i \times p)$, $\boldsymbol{\alpha}$, $(p \times 1)$, $\boldsymbol{Z}_i$, $(n_i \times k)$, $\boldsymbol{b}_i$, $(k \times 1)$ and $\boldsymbol{e}_i$, $(n_i \times 1)$, Stage 1 is for the individual subjects, i,

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\alpha} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i \,, \tag{7}$$

such that $\boldsymbol{e}_i \sim N(\boldsymbol{0}, \boldsymbol{R}_i)$ with $\boldsymbol{R}_i$ $(n_i \times n_i)$. At this stage, $\boldsymbol{\alpha}$ and $\boldsymbol{b}_i$ are assumed to be fixed, and the $\boldsymbol{e}_i$ are assumed to be independent. Stage 2 is for the population. The assumption is that $\boldsymbol{b}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$ independently of one another and $\boldsymbol{e}_i$, such that $\boldsymbol{D}$ is $(k \times k)$. Only the $\boldsymbol{\alpha}$ are assumed to be fixed at this stage. Marginally, $\boldsymbol{y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\alpha}, \boldsymbol{R}_i + \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i')$.

The model has several nice statistical features: 1) no requirement for balanced data, 2) explicit modeling and analysis of between- and within-subject variation, and 3) the

capability of handling missing data without eliminating all values for a particular subject. The multivariate model commonly used in practice at the time did not allow for the definition and estimation of random individual characteristics. Furthermore, mixed model techniques allowing unbalanced or incomplete repeated measures offered a great deal more flexibility than the strict assumptions required in MULTIREP practices. However, very little work towards accurate inference in mixed model theory existed. In this respect, as long as there was not a need to specify a covariance structure, MULTIREP and UNIREP analyses stood, and still stand, above mixed models, especially in cases of complete data.

Although practical mixed model analyses were now available, Koele (1982) acknowledged that the power methods for mixed models available at the time did not perform well. He did not offer a solution to this problem.

Jennrich and Schluchter (1986) presented techniques that took advantage of specific covariance forms, such as compound symmetry (CS) and first order autoregressive (AR1). They also illustrated how, based on the design of the model, one could allow groups to have different covariance matrices from others. Other commonly used mixed model techniques discussed in their paper included Newton-Raphson and Fisher scoring algorithms, maximum likelihood (ML) and restricted maximum likelihood (REML) methods. These techniques are also discussed in Laird and Ware (1982).

Catellier and Muller (2000) evaluated the effectiveness of UNIREP, MULTIREP and mixed models on inference techniques in cases with data missing at random and missing completely at random. They observed simulated test sizes as high as 0.59 with a target of 0.05 for the mixed model test of the interaction between the repeated measure and the grouping factor with complete and balanced data. Meanwhile, the UNIREP and MULTIREP tests controlled the type I error rate at or below the nominal rate. Catellier and Muller (2000) further noted that, while mean estimates often coincide between mixed and multivariate models, hypothesis testing and confidence intervals usually differ greatly.

They recommended using the UNIREP and MULTIREP techniques over those of mixed models whenever appropriate.

Gueorguieva and Krystal (2004) fully supported the use of the mixed model over UNIREP and MULTIREP methods, however. Although, their reasons were more directed towards convenience. When applied to psychiatric studies frequently with mistimed or missing data for many subjects, they claimed that the appeal for use of the mixed model was due to flexibility of use. They also cited the prevalence and availability of mixed model software as a reason. All of their examples had large sample sizes, which allowed for more accurate asymptotic approximations. Gueorguieva and Krystal (2004) did not mention inference accuracy in detail. However, they did acknowledge that small samples may bias parameter estimates and statistical tests in mixed models. They claimed that in their field of psychiatry, larger sample sizes and missing data are common.

Muller and Stewart (2006) demonstrated that MULTIREP and UNIREP are special cases of mixed models. An oversimplification of their explanation is that MULTIREP techniques require $\boldsymbol{Z}_i = \mathbf{0}$ with an unstructured covariance matrix, $\boldsymbol{R}_i$. Traditionally, UNIREP techniques require $\boldsymbol{Z}_i = \mathbf{0}$ with a spherical or near spherical covariance matrix required.

1.6 A History of UNIREP Tests

This section contains a review of work performed towards the development of UNIREP tests. The reviewed papers provide the basic theory behind the methods used and the theory developed in Chapters 2-4 of this paper. Also, the reasons for the development of the Huynh-Feldt sphericity estimator are discussed. More relevant to the current research, the need for a more general Huynh-Feldt sphericity estimator is introduced.

Box (1954a, b) described consequences of violating the assumption of homogeneity within subjects using the uncorrected UNIREP test. He began by noting that any real quadratic form of rank $r \leq p$, $Q = \boldsymbol{z}'\boldsymbol{A}\boldsymbol{z}$, such that $\boldsymbol{z} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, is distributed as a

weighted finite sum of chi-square random variables. In particular, $Q = \sum_{j=1}^{r} \lambda_j \chi^2(1)$, such that each χ^2 is independent and the λ_j's are the r non-zero eigenvalues of $\boldsymbol{\Sigma A}$. From this, he was able to calculate exactly a ratio of quadratic forms,

$$T = Q_1/Q_2 = \left[\sum_{j=1}^{r'} \lambda'_j \chi^2(\nu'_{j'})\right] \Big/ \left[\sum_{j=1}^{r} \lambda_j \chi^2(\nu_j)\right]. \tag{8}$$

He showed that the UNIREP test statistic, $T_u = [\text{tr}(\widehat{\boldsymbol{\Delta}})/a]/[\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)]$, could be approximated with an F distribution under the null hypothesis, $F(ab\epsilon, ab\nu_e\epsilon)$, even if sphericity was not met.

Geisser and Greenhouse (1958) offered an extension to Box (1954a, b) for UNIREP test statistics. They provided bounds for ϵ, which are independent of the elements of the covariance matrix, $1/b \leq \epsilon \leq 1$, such that $b = \text{rank}(\boldsymbol{U})$. When sphericity came into question, they recommended either using the lower bound as an estimate of ϵ, thus $T_u \sim F(a, a\nu_e)$, or the maximum likelihood estimator (MLE),

$$\widehat{\epsilon} = \frac{\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)}{b\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)}\,, \tag{9}$$

now known as the Geisser-Greenhouse estimator. Use of the former, known as the Box conservative estimator, yields conservative results.

Greenhouse and Geisser (1959) acknowledged that when sphericity was not met, most had approached the problem using MULTIREP techniques. However, they recommended using approximate F distributions instead, $F(ab\epsilon, ab\nu_e\epsilon)$, because they were easier to compute than the MULTIREP statistics. Based on the bounds on ϵ presented by Geisser and Greenhouse (1958), clearly ϵ reduces degrees of freedom for the approximate tests. The measure of sphericity, ϵ, is a function of the population covariance matrix. However, the population covariance matrix is rarely known and is instead estimated from the sample.

Greenhouse and Geisser (1959) suggested using the conservative test offered in their 1958 paper, unless the covariance matrix is estimated with a large number of degrees of freedom.

Cole and Grizzle (1966) discussed known methods to test the hypothesis $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$, thus allowing one to test sphericity. Furthermore, they confirmed the work and conclusions offered by Geisser and Greenhouse (1958, 1959). Cole and Grizzle (1966) also compared the use of MULTIREP and UNIREP tests when sphericity was not met, or not known to be met. When they wrote their paper, only the uncorrected and Box conservative UNIREP tests were used in common practice. Cole and Grizzle (1966) noted that there is some loss of power when using MULTIREP techniques, and that the loss of power is most obvious for the tests for the single degree of freedom contrasts. The reasoning behind such a loss in power begins with the realization that each test essentially consists of a comparison between a single squared deviation and an estimate of a variance. The assumptions required for the UNIREP analysis allow this estimate to be based on more degrees of freedom than its counterpart. Cole and Grizzle (1966) did point out that in the case of $a = 1$, each MULTIREP test is equivalent, and in the case of $b = 1$, UNIREP and MULTIREP tests are equivalent.

Huynh and Feldt (1970) discussed the necessary conditions for the UNIREP tests to be distributed F. Previous work had already shown that if the outcomes were normally distributed and the covariance was compound symmetric, then the mean square ratios for the treatment, group and treatment by group interaction followed exact F distributions in a two-way ANOVA. The distributions are central if the null hypothesis is true. Huynh and Feldt (1970) showed that these ratios were distributed F under more general conditions for which compound symmetry represented a specific case. The condition they presented as necessary was called sphericity. This condition is met if and only if the covariance matrix of $\boldsymbol{Y}_* = \boldsymbol{U}\boldsymbol{Y}$ is $\boldsymbol{\Sigma}_* = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}' = \lambda \boldsymbol{I}_b$, i.e. proportional to an identity matrix. More generally, stating that $\boldsymbol{\Sigma}$ satisfies sphericity for a comparison with b degrees of freedom

amounts to putting $[b(b+1)/2]-1$ linear constraints among the variances and covariances of $\mathbf{\Sigma}$. They suggested using Mauchly's sphericity test to check this assumption, a strategy later determined to be a poor approach.

Rouanet and Lepine (1970) compared MULTIREP and UNIREP tests, and, similar to many authors before them, advocated UNIREP tests because they saw them as generally more powerful. However, this is not always the case because, for $\min(p, s) > 1$ without sphericity being met, there is no uniformly most powerful test. In this case, the choice of test depends on the alternative and the degree to which sphericity is not met. Rouanet and Lepine (1970) also examined the requirements for use of UNIREP tests. They agreed with Huynh and Feldt (1970) by claiming that compound symmetry was merely a sufficient requirement, and that sphericity was the necessary requirement. Rouanet and Lepine (1970) also recommended Mauchly's sphericity test to check this assumption, but warned that tests about variances and covariances are known to be sensitive to nonnormality. They suggested that normality should be checked.

In his 1972 paper, Davidson offered his thoughts on choices among MULTIREP and UNIREP tests. Still, only evaluating the uncorrected and Box conservative UNIREP tests, Davidson (1972) recommended UNIREP tests over MULTIREP for small sample sizes, due to the fact that the latter were much less powerful in the cases he considered. However, for large sample sizes he advocated MULTIREP tests. He reasoned that the standard test of sphericity had acceptable power only when the MULTIREP tests of the hypothesis was essentially as powerful as the uncorrected UNIREP. When one does not concern oneself with checking sphericity, practice suggests use of the Box conservative test, which he showed to perform from somewhat better to much worse than the MULTIREP tests. The choice between the Box conservative and the MULTIREP tests depended mostly on the extent to which the covariance matrix was spherical. Of the tests considered in his paper, he concluded that if the covariance was not known ahead of time, only the Box conservative

and the MULTIREP tests allowed the researcher to control the type I error. Today, testing sphericity with Mauchly's test is not necessary due to the abilities of the Geisser-Greenhouse and Huynh-Feldt tests to control test size despite covariance structure.

By the mid 1970's, the Box conservative test, using $\epsilon = 1/b$, and the Geisser-Greenhouse test, using $\widehat{\epsilon}$, were both used in practice. However, Huynh and Feldt (1976) and Huynh (1978) cited examples where $\widehat{\epsilon}$ might be seriously biased if the population sphericity was approximately 0.75. This bias resulted in an overcorrection of the degrees of freedom and implied a more stringent significance level than the nominal level desired. Huynh and Feldt (1976) responded to this problem by introducing a new estimator for sphericity.

Under the assumption of multivariate normality, $\widehat{\epsilon}$ is the MLE for ϵ. The MLE is biased when the population is homogeneous. By calculating expected values of the numerator and denominator of the ratio $\epsilon = [\text{tr}^2(\boldsymbol{\Sigma}_*)]/[b\text{tr}(\boldsymbol{\Sigma}_*^2)]$, Huynh and Feldt (1976) developed, what they claimed to be, a ratio of unbiased estimators, $\widetilde{\epsilon} = (Nb\widehat{\epsilon} - 2)/[b(\nu_e - b\widehat{\epsilon})]$. Huynh and Feldt (1976) determined that $\widetilde{\epsilon} \geq \widehat{\epsilon}$ with equality if $\widehat{\epsilon} = 1/b$. The difference between the two estimates decreased as sample size increased. Huynh and Feldt (1976) further noted that their estimate occasionally exceeded 1, and should be truncated to 1.0 in such cases. Their simulations showed that while $\widehat{\epsilon}$ was a less biased estimator than $\widetilde{\epsilon}$ when $\epsilon \leq 0.5, \widetilde{\epsilon}$ was less biased when $\epsilon > 0.75$.

In Chapter 2, the Huynh-Feldt estimator is shown to be a ratio of unbiased estimators only for the special case of rank of $\boldsymbol{X}$ equal to 1. Huynh and Feldt (1976) did not derive an estimator that may be used for any rank of $\boldsymbol{X}$. As a result, the Huynh-Feldt test and power calculation may be biased when rank of $\boldsymbol{X}$ is greater than 1.

Huynh and Mandeville (1979) reviewed past works on the condition of sphericity citing Huynh and Feldt (1970) and Rouanet and Lepine (1970) as the authors who showed that compound symmetry is a sufficient property, but not necessary. In general, Huynh and

Mandeville (1979) concurred with the need for sphericity to be met and noted that this condition is not based on the orthonormal variables or on the repeated measures.

Wallenstein and Fleiss (1979) took a unique approach to specifying a lower limit for ϵ by considering specific covariance structures, such as AR1. Geisser and Greenhouse (1958) showed a lower bound for ϵ to be $1/b$, for the general case. Wallenstein and Fleiss (1979) showed that when the covariance structure is AR1, the $\min(\epsilon) = \lim_{\rho \to 1} \epsilon =$ $[5(p+1)]/(2p^2+7) \geq 1/b$ when $p > 2$ with equality at $p = 2$. Here, p is defined as the number of responses per subject and $b = \text{rank}(\boldsymbol{U}) = (p-1)$ for their examples. They further showed that this bound applies to the following covariance structures as well:

1) $\boldsymbol{\Sigma} = \sigma_b^2(\mathbf{1}'\mathbf{1}) + \sigma_e^2\boldsymbol{S}$, such that $\boldsymbol{S}$ is AR1 and σ_b^2 represents the subject effect variance with the subject effect assumed to be $N(0, \sigma_b^2)$,

2) $\boldsymbol{\Sigma} = \sigma_b^2(\boldsymbol{S}) + \sigma_e^2\boldsymbol{I}$, such that $\boldsymbol{S}$ is AR1 and σ_b^2 represents the subject effect variance with the subject effect assumed to be $N(0, \sigma_b^2)$. The latter is more appropriate in cases in which time points are not as close to one another.

Huynh and Feldt (1980) took a closer look at the theoretical derivation of the UNIREP F tests, and considered the ramifications for various assumption violations. They noted that the test of interaction is more vulnerable to conditions of covariance heterogeneity than the tests for main effects. Also, they observed that the traditional F test in repeated measures designs with identical covariance matrices will err on the liberal side (i.e. show a size larger than the nominal test size), especially when ϵ and N are small. They further gave examples of how high correlation results in smaller residual error, and thus greater power for the test when sphericity was not met.

O'Brien and Kaiser (1985) suggested MULTIREP over the uncorrected UNIREP test in nearly all cases. They claimed that repeated measures are rarely independent and that the conditions implied as necessary by the uncorrected UNIREP test are too severe. Additionally, they claimed that pretesting with Mauchly's sphericity test had the following

shortcomings: 1) if there was insufficient sample size, sphericity may be accepted, even if not warranted, and 2) Mauchly's test was very sensitive to violations of normality. Specifically, they claimed that Mauchly's test tended to accept sphericity too often for light tailed distributions and rejected sphericity for heavy tailed distributions. Huynh and Mandeville (1979) showed that these tendencies were amplified by increasing sample size. O'Brien and Kaiser (1985) believed that so much work was required for testing sphericity, that simply moving to the MULTIREP tests was more logical, despite the lost power. Today, there is no need to test sphericity because the Geisser-Greenhouse and Huynh-Feldt tests are capable of controlling test size, even when sphericity is not met.

O'Brien and Kaiser (1985) evaluated the results of Davidson (1972) and Huynh (1978), among others, who had compared the power of the Box conservative UNIREP test to MULTIREP tests. Overall, O'Brien and Kaiser (1985) found that no procedure was always, or even usually, the most powerful.

Catellier and Muller (2000) developed hypothesis tests for Gaussian repeated measures with missing data, accurate in small samples. Along with describing analogs of several MULTIREP tests, they developed techniques for the Geisser-Greenhouse test. When compared to the now popularized mixed model techniques, they showed that for small samples, even with no missing data, the mixed model had inflated type I error rates. Meanwhile, the UNIREP and MULTIREP tests controlled the type I error rates at or below the nominal rate. Thus, the approximate F tests were essentially unbiased for complete data.

Coffey and Muller (2003) extended their work on the general linear univariate model for internal pilots to UNIREP, providing mostly approximate results. They indicated that UNIREP required complete and consistently timed data within-subject and did not allow for repeated covariates. Like Catellier and Muller (2000), Coffey and Muller (2003) preferred UNIREP over mixed models, when applicable, due to superior control of test size,

especially in small samples. They also noted that UNIREP power approximations, using the Muller and Barton (1989) approximation (discussed in section 1.7), had had extensive study, while power approximations for mixed models had not.

1.7 Power for UNIREP Tests

This section contains a review of work performed towards the development of power calculations for UNIREP tests. Specifically, the power approximations for UNIREP tests developed by Muller and Barton (1989), and later improved upon by Muller *et al.* (2007), provide much of the background and distributional approximations needed for the theoretical development of confidence intervals for power for UNIREP tests presented in Chapter 4. Additional background theory is reviewed in section 1.8: Confidence Intervals for Power.

Boik (1981) offered some basic work on power for UNIREP tests under nonsphericity. He demonstrated that even small departures from sphericity could result in serious changes to test size and power. Boik (1981) cited various studies, including Huynh (1978), that had shown that the Geisser-Greenhouse test was slightly negatively biased (i.e. test size is less than α), with the greatest bias with minimal departures from sphericity. For these minimal departures, he suggested using Huynh-Feldt as a test that produces a test size closer to the nominal α level.

Muller and Barton (1989) offered more on the topic of power for UNIREP tests than anyone before them. They provided power equations for all four UNIREP tests: 1) the uncorrected (Box, 1954a, b), 2) the Huynh-Feldt (1976), 3) the Geisser-Greenhouse (1958, 1959), and 4) the Box conservative (Geisser and Greenhouse, 1958), with sphericity multipliers

$$\begin{array}{ccccccc} \text{Box} & & \text{GG} & & \text{HF} & & \text{UN}. \\ 1/b & \leq & \widehat{\epsilon} & \leq & \widetilde{\epsilon} & \leq & 1 \end{array}$$

Muller and Barton (1989) noted that the UNIREP approach allows for fewer subjects with equal power, with sphericity met, when compared to the MULTIREP approach. Muller and Peterson (1984) had provided accurate power approximations for Wilks, Pillai-Bartlett and Hotelling-Lawley tests.

As mentioned previously, the corrected tests decrease the degrees of freedom of the approximate F distribution, and thus increase the critical value, leading to decreased power. The order of increasing critical value, decreasing test size (type I error rate) and decreasing power is UN, HF, GG, Box. In order to calculate power for all UNIREP tests using the Muller-Barton approach, there is a need for a noncentral F distribution that can handle fractional degrees of freedom.

Muller and Barton (1989) found that the agreement of the power calculations was excellent with the biggest differences at $\epsilon = 1$ for the corrected tests, usually in small samples. Obviously, the exact uncorrected test could be used here. Without prior knowledge of the sphericity, Muller and Barton (1989) suggested using the Geisser-Greenhouse test because their simulations and examples showed acceptable type I error control and maximization of power.

Muller and Benignus (1992) provided an informative review of the power methods for univariate analysis, laying the groundwork for future developments in power for UNIREP tests. Muller *et al.* (1992) evaluated the multivariate power techniques given by Muller and Barton (1989) and Muller and Peterson (1984) for UNIREP and MULTIREP, respectively. They emphasized the importance of power analysis in choosing a study design and testing strategy. They described two possible mistakes that might occur if the power analysis and data analysis were misaligned. First, choosing a sample size too small would lead to a study with inadequate sensitivity, and second, choosing a sample size too large would lead to wasted resources.

Muller and Peterson (1984) claimed that in order to calculate power for the MULTIREP tests, only α, $\boldsymbol{\Sigma}$, $\boldsymbol{X}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{U}$ and $\boldsymbol{\Theta}_0$ need to be specified. The method produces exact results for $s = \min(a, b) = 1$. Power calculations for UNIREP tests are quite similar. If sphericity holds, the uncorrected UNIREP power is exact. For the corrected UNIREP tests, the power calculations are a simple extension, but approximate. Muller *et al.* (1992) showed that often there is no uniformly most powerful test, so the choice of test depends on the alternative and the degree to which sphericity is not met.

Muller *et al.* (2007) discussed the advantages of UNIREP methods and disadvantages of mixed models in relation to imaging studies, focusing on how UNIREP controls test size and offers better power methods than mixed models. They suggested always using UNIREP or MULTIREP over mixed models, when applicable. "Many applications with correlated outcomes in medical imaging and other fields have simple properties which do not require the generality of a mixed model." In imaging studies, there are often repeated measures with no missing or mistimed data and small sample sizes.

More relevant to this research, Muller *et al.* (2007) created a better UNIREP power approximation. Coffey and Muller (2003) gave cases where the Muller and Barton (1989) approximations failed to provide even one digit of accuracy for the Geisser-Greenhouse test. Using the noncentral distribution function approximation presented by Kim *et al.* (2006), Muller *et al.* (2007) were able to give second order approximations for the uncorrected and Box conservative tests, yielding exact power. They also provided approximate power for the Geisser-Greenhouse and Huynh-Feldt tests by combining the CDF approximation with approximations for the expected degrees of freedom. They showed that the test statistic could be approximated by a noncentral F distribution, such that

$$\Pr\{T_u \leq t\} \approx \Pr\left\{ \frac{\lambda_1 y_1/(ab)}{\lambda_2 y_2/(b\nu_e)} \leq t \right\} = F_F(t; \nu_1, \nu_2, \omega) \,, \tag{10}$$

with $y_1 \sim \chi^2(\nu_1, \omega)$, $y_2 \sim \chi^2(\nu_2)$, $\text{tr}(\widehat{\boldsymbol{\Delta}}) \approx \lambda_1 y_1$ and $\text{tr}(\widehat{\boldsymbol{\Sigma}}_*) \approx \lambda_2 y_2$. For the sake of clarity of presentation, the right side of equation 10 is shown simplified by way of Lemma 4.2 in section 4.3 of this paper. Through simulations, they demonstrated that the new power approximations eliminated most inaccuracies in existing methods for all four UNIREP tests.

1.8 Confidence Intervals for Power

When designing a study, accurate power analysis is essential to enhancing study design efficiency. Often the variance is estimated in this analysis, and power becomes a random variable. Providing confidence intervals for these random power values would be useful in any study design. A lower bound for power would allow stating that a study has power of at least "P" to detect an effect, with a specified confidence.

This section contains a review of work performed towards the development of power confidence intervals. Much of the theoretical background needed for the development of confidence intervals for power for UNIREP tests presented in Chapter 4 come from three papers. Taylor and Muller (1995, discussed below) developed exact power confidence intervals for estimated variance and known means in the univariate case. Muller and Barton (1989) and Muller *et al.* (2007) provided the methods for power calculations in the UNIREP setting. These methods are discussed in section 1.7.

Dudewicz (1972) was one of the first to discuss methods for confidence intervals for power in a univariate setting. He suggested substituting approximate confidence bounds for σ^2 into power calculations for a t test. This approach resulted in approximate confidence limits for power. Although he did not present any asymptotic or simulation evidence as to the accuracy of his method, he claimed the technique was quite good, especially for small samples. Venables (1975) discussed confidence intervals for noncentrality in noncentral chi-square and F distributions.

Taylor and Muller (1995) developed exact confidence intervals for power of the univariate linear model with estimated variance and fixed mean. They began in much the same way as Dudewicz (1972), with an estimated variance, which implies that power for a fixed sample size must be recognized as a random variable. The univariate test statistic is

$$F_{obs} = \frac{S_H(\widehat{\boldsymbol{\theta}},\, N)/a}{S_E/\nu_e}\ , \tag{11}$$

such that S_H is a function of sample size through $\boldsymbol{X}$. Gaussian theory leads to

$$\frac{\nu_e\widehat{\sigma}^2}{\sigma^2} \sim \chi^2(\nu_e)\ . \tag{12}$$

Thus, a confidence interval for σ^2 is found like so:

$$\Pr\left\{ \frac{\widehat{\sigma}^2\nu_e}{c_{crit}(1-\alpha_{cU}|\nu_e)} < \sigma^2 < \frac{\widehat{\sigma}^2\nu_e}{c_{crit}(\alpha_{cL}|\nu_e)} \right\} = 1-\alpha_L-\alpha_U\ , \tag{13}$$

such that $(1-\alpha_{cL}-\alpha_{cU})$ is the confidence coefficient and $c_{crit}(\alpha_{cL}|\nu_e) = \alpha_{cL}$ quantile for central $\chi^2(\nu_e)$. Similarly for $1-\alpha_{cU}$. In the univariate case, the expression for noncentrality is $\omega = (S_H/\sigma^2)$. Thus, the confidence interval for noncentrality is

$$\Pr\left\{ \frac{(S_H)\ c_{crit}(\alpha_{cL}|\nu_e)}{\widehat{\sigma}^2\nu_e} < \omega < \frac{(S_H)\ c_{crit}(1-\alpha_{cU}|\nu_e)}{\widehat{\sigma}^2\nu_e} \right\} = 1-\alpha_L-\alpha_U\ , \tag{14}$$

such that $\widehat{\omega}_L = c_{crit}(\alpha_{cL}|\nu_{2e}) \cdot [S_H(\boldsymbol{\theta},\, N)/S_E]$ and
$\widehat{\omega}_U = c_{crit}(1-\alpha_{cU}|\nu_{2e}) \cdot [S_H(\boldsymbol{\theta},\, N)/S_E]$, and $S_E = \widehat{\sigma}^2\nu_e$. These bounds provide exact confidence intervals for ω, such that $0 \le \omega \le \infty$.

Due to the strict monotone dependence of the noncentral F distribution function on noncentrality, an exact confidence interval for power follows from an exact confidence interval for ω. Thus, exact power confidence limits are

$$\widehat{P}_L = 1 - F_F[f_{crit}(1-\alpha_t)|a,\, \nu_e, \widehat{\omega}_L] \tag{15}$$

and

$$\widehat{P}_U = 1 - F_F[f_{crit}(1-\alpha_t)|a,\, \nu_e, \widehat{\omega}_U]\,. \tag{16}$$

Taylor and Muller (1995) also demonstrated how to construct one-sided confidence intervals for power using the same method. Additionally, they broadened their technique to allow for the development of exact confidence regions for the whole of the power curve. Muller and Fetterman (2002) described how to use existing power software to compute exact power confidence intervals in univariate analyses, as presented by Taylor and Muller (1995).

Taylor and Muller (1996) extended their 1995 paper by discussing confidence intervals for univariate power when both the variance and mean are estimated. They described how power calculations may result in biased estimators, but unbiased bounds. No research has been published with respect to power confidence intervals for UNIREP, despite the commonness of such designs in practice.

Chapter 2
A More General Version of the Huynh-Feldt Sphericity Estimator

2.1 Motivation

When sphericity is not met, Box (1954a, b) showed that the UNIREP test statistic under the null could be approximately distributed as an F with reduced degrees of freedom, $F(ab\epsilon, b\nu_e\epsilon)$. The reduction comes from various estimators of ϵ, a measure of sphericity that ranges from $1/b$ to 1, which are used as degrees of freedom multipliers. Geisser and Greenhouse (1958) offered the Box conservative estimate, $\epsilon = 1/b$, and the maximum likelihood estimator, $\widehat{\epsilon} = [\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)]/[b\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)]$, now known as the Geisser-Greenhouse estimator, as degrees of freedom multipliers.

In their 1976 paper, Huynh and Feldt described examples of when the Geisser-Greenhouse estimator was seriously biased, most often in nearly spherical populations. They found that, in such cases, the estimator overcorrected the degrees of freedom, resulting in a more stringent significance level than the nominal level desired. Huynh and Feldt (1976) responded with an estimator of ϵ which they showed through examples to be less biased and less dependent on large sample sizes than the Geisser-Greenhouse estimator when the population covariance deviated only moderately from sphericity. They claimed that their estimator, $\widetilde{\epsilon} = (Nb\widehat{\epsilon} - 2)/[b(\nu_e - b\widehat{\epsilon})]$, was a ratio of unbiased estimators, and showed that $\widetilde{\epsilon} \geq \widehat{\epsilon}$ with equality if $\widehat{\epsilon} = 1/b$. However, their claim holds true only if rank of $\boldsymbol{X}$ is equal to 1, resulting in $\nu_e = N - 1$ error degrees of freedom between subjects. They further illustrated through simulations that, while $\widehat{\epsilon}$ is a better estimator of the population sphericity parameter than $\widetilde{\epsilon}$ when $\epsilon \leq 0.5$, $\widetilde{\epsilon}$ is less biased when $\epsilon > 0.75$.

The Huynh-Feldt estimator will be shown to be a ratio of unbiased estimators only for the special case of rank of $\boldsymbol{X}$ equal to 1. As a result, the Huynh-Feldt test and power calculation may be biased in cases with rank of $\boldsymbol{X}$ greater than 1. An estimator composed of a ratio of unbiased estimators for any rank of $\boldsymbol{X}$ will be presented and evaluated for a wide range of simulations. For the sake of clarity, the Huynh-Feldt estimator, $\widetilde{\epsilon}$, will hence forth be referred to as $\widetilde{\epsilon}_{HF}$.

2.2 Notation and Known Results

The estimated covariance matrix among response variables is $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{Y}'[\boldsymbol{I}_N - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}']\boldsymbol{Y}/\nu_e$, and the estimated covariance matrix among transformed (hypothesis) variables is $\widehat{\boldsymbol{\Sigma}}_* = \boldsymbol{U}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{U}$. Also, $\boldsymbol{S} = \nu_e\widehat{\boldsymbol{\Sigma}}_* \sim \mathcal{W}_b(\nu_e, \boldsymbol{\Sigma}_*, \boldsymbol{0})$, with $\boldsymbol{\lambda}$, $(b \times 1)$, the eigenvalues of $\boldsymbol{\Sigma}_*$. If $\tau_1 = \text{tr}^2(\boldsymbol{\Sigma}_*) = (\boldsymbol{1}_b'\boldsymbol{\lambda})^2$ and $\tau_2 = \text{tr}(\boldsymbol{\Sigma}_*^2) = \boldsymbol{\lambda}'\boldsymbol{\lambda}$, then the population sphericity parameter can be written as a ratio of two parameters, $\epsilon = [\text{tr}^2(\boldsymbol{\Sigma}_*)]/[b\text{tr}(\boldsymbol{\Sigma}_*^2)] = \tau_1/(b\tau_2)$. If $\widetilde{\tau}_1 = \text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) = \text{tr}^2(\boldsymbol{S})/\nu_e^2$ and $\widetilde{\tau}_2 = \text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) = \text{tr}(\boldsymbol{S}^2)/\nu_e^2$, maximum likelihood estimation gives the Geisser-Greenhouse estimator as $\widehat{\epsilon} = [\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)]/[b\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)] = \widetilde{\tau}_1/(b\widetilde{\tau}_2)$. Note that while $\widehat{\boldsymbol{\Sigma}}_*$ is an unbiased estimator of $\boldsymbol{\Sigma}_*$, $\widetilde{\tau}_1$ and $\widetilde{\tau}_2$ are not unbiased estimators of τ_1 and τ_2, respectively. For $t_1 = \text{tr}^2(\boldsymbol{S}) = \nu_e^2\widetilde{\tau}_1$, Muller *et al.* (2007, Appendix A) proved

$$\text{E}\big[\text{tr}^2(\boldsymbol{S})|\nu_e \geq b\big] = 2\nu_e\text{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e^2\text{tr}^2(\boldsymbol{\Sigma}_*) \tag{17}$$

and, for $t_2 = \text{tr}(\boldsymbol{S}^2) = \nu_e^2\widetilde{\tau}_2$,

$$\text{E}\big[\text{tr}(\boldsymbol{S}^2)|\nu_e \geq b\big] = \nu_e(\nu_e + 1)\text{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e\text{tr}^2(\boldsymbol{\Sigma}_*)\ . \tag{18}$$

2.3 A Rank-Adjusted Huynh-Feldt Sphericity Estimator

Based on the moments presented in the previous section, unbiased estimators for both τ_1 and τ_2 may be derived. The unbiased estimators are functions of the biased estimators $\widetilde{\tau}_1$ and $\widetilde{\tau}_2$, and are introduced in Lemma 2.1 below. In Lemma 2.2, an approximately unbiased

sphericity estimator, $\widetilde{\epsilon}_r$, is described for any rank of $\boldsymbol{X}$ as the ratio of the two unbiased, but correlated, estimators for τ_1 and τ_2.

Lemma 2.1 Unbiased estimators for $\tau_1 = \text{tr}^2(\boldsymbol{\Sigma}_*)$ and $\tau_2 = \text{tr}(\boldsymbol{\Sigma}_*^2)$ are

$$\begin{aligned} \widehat{\tau}_1 &= [t_1 - 2(\nu_e + 1)^{-1} t_2] \nu_e^{-2} \{1 - 2[\nu_e(\nu_e + 1)]^{-1}\}^{-1} \\ &= [\widetilde{\tau}_1 - 2(\nu_e + 1)^{-1} \widetilde{\tau}_2] \{1 - 2[\nu_e(\nu_e + 1)]^{-1}\}^{-1} \end{aligned} \tag{19}$$

$$\begin{aligned} \widehat{\tau}_2 &= (\nu_e t_2 - t_1) \{\nu_e[\nu_e(\nu_e + 1) - 2]\}^{-1} \\ &= \left(\nu_e^2 \widetilde{\tau}_2 - \nu_e \widetilde{\tau}_1\right) [\nu_e(\nu_e + 1) - 2]^{-1} . \end{aligned} \tag{20}$$

Lemma 2.2 A ratio estimating ϵ in terms of correlated, but unbiased, estimators is

$$\begin{aligned} \widetilde{\epsilon}_r &= \frac{\widehat{\tau}_1}{b\widehat{\tau}_2} \\ &= \frac{(\nu_e + 1) b\widehat{\epsilon} - 2}{b(\nu_e - b\widehat{\epsilon})} . \end{aligned} \tag{21}$$

Corollary 2.3 If $\text{rank}(\boldsymbol{X}) = 1$, then $\widetilde{\epsilon}_r = \widetilde{\epsilon}_{HF}$, the estimator proposed by Huynh and Feldt (1976).

Huynh and Feldt (1976) noted cases in which their estimator exceeded a value of 1.0. In turn, they suggested using a truncated estimator, $\min(\widetilde{\epsilon}_{HF}, 1)$. Theirs is a special case of the newly proposed rank-adjusted estimator, so a similar truncation must be performed. Thus, an approximately unbiased estimator derived from a ratio of unbiased estimators for any rank of $\boldsymbol{X}$ is

$$\widetilde{\epsilon}_r = \min\left[\frac{(\nu_e + 1) b\widehat{\epsilon} - 2}{b(\nu_e - b\widehat{\epsilon})}, 1\right] . \tag{22}$$

Obviously, $\widehat{\epsilon} \leq \widetilde{\epsilon}_r \leq \widetilde{\epsilon}_{HF}$, with equality of the three estimators if and only if $\widehat{\epsilon} = 1/b$ and rank of $\boldsymbol{X}$ is equal to 1. Equality of $\widehat{\epsilon}$ and $\widetilde{\epsilon}_r$ exists if $\widehat{\epsilon} = 1/b$, and equality of $\widetilde{\epsilon}_r$ and $\widetilde{\epsilon}_{HF}$ exists if rank of $\boldsymbol{X}$ is equal to 1.

2.4 Simulations

The new rank-adjusted approximately unbiased sphericity estimator was evaluated for a wide range of simulations. All of the simulations considered the condition of rank of $\boldsymbol{X}$ greater than 1. The simulated realizations consisted of $p = 5$ repeated measures, N the sample size of 16, 32 and 48, and q the rank of $\boldsymbol{X}$ equal to 4, 8 and 16, in the model

$$\underset{(N \times 5)}{\boldsymbol{Y}} = \underset{(N \times q \times 5)}{\boldsymbol{XB}} + \underset{(N \times 5)}{\boldsymbol{E}} . \tag{23}$$

Population covariance matrices were chosen to provide specific population sphericity values, $\epsilon \in \{0.282, 0.505, 0.720, 1.00\}$. Specific design matrices, $\boldsymbol{X}$, were defined. Matrices of regression coefficients, $\boldsymbol{B}$, were defined as $\mathbf{0}$, $(q \times 5)$, to illustrate the null case.

Pseudo-random realizations of the error matrix, $\boldsymbol{E}$, were generated, and both the Huynh-Feldt and new rank-adjusted sphericity estimates were calculated and tabulated for 500,000 replications per condition. This number of replications was chosen to ensure a standard error of observed mean estimates less than or equal to 0.0003, nearly guaranteeing 3 digits of accuracy. The calculated sphericity estimates were then compared to the population sphericity values. Appendix C contains a more detailed description of the simulation parameters. All simulations were conducted in SAS/IML (SAS 9.1, SAS Institute, 2003). Software that performs a wide variety of General Linear Multivariate Model computations called LINMOD 3.3 (http://ehpr.ufl.edu/muller/) was modified to include the rank-adjusted estimator and test. The modified version was used in all simulations and will be made available soon.

In Table 2.1, the mean and maximum absolute deviations of the Huynh-Feldt and rank-adjusted sphericity estimates from the population sphericity values are compared. These deviations have been averaged over all sample sizes and ranks of $\boldsymbol{X}$ considered for three of the four predetermined population sphericity values spanning the range of possible values.

The out performance of the Huynh-Feldt sphericity estimate by the rank-adjusted in terms of accuracy is immediately evident, for each of the population sphericity values.

Table 2.1. Mean / (Max) Absolute Deviations ($Observed - Population$) of HF and Rank-Adjusted Sphericity Estimates averaged over N and Rank(X), Standard Error of Observed $\leq$ 0.0003.

	HF	Rank-Adjusted
ϵ	$\widetilde{\epsilon}$	$\widetilde{\epsilon}_r$
0.282	0.131 / (0.314)	0.008 / (0.022)
0.505	0.229 / (0.447)	0.049 / (0.111)
0.720	0.189 / (0.279)	0.028 / (0.059)

In Tables 2.2-2.5, mean deviations between the Huynh-Feldt, rank-adjusted, and Geisser-Greenhouse sphericity estimates and the four predetermined population sphericity values are presented for the various sample sizes and ranks of $\boldsymbol{X}$ considered. For every case in Table 2.2 with population sphericity of 0.282, the rank-adjusted sphericity estimates better estimated the population sphericity value than the Huynh-Feldt estimates. The accuracy of the rank-adjusted estimates compared to those of the Huynh-Feldt seems to have improved as the sample size and rank of $\boldsymbol{X}$ increased. For the smallest sample size and smallest rank of $\boldsymbol{X}$ considered, $N = 16$ and rank$(\boldsymbol{X}) = 4$, the deviation between the Huynh-Feldt estimate and the population sphericity value was approximately 7 times that of the deviation between the rank-adjusted estimate and the population sphericity value. This difference in magnitude increased to approximately 33 for the largest sample size and the largest rank of $\boldsymbol{X}$ considered, $N = 48$ and rank$(\boldsymbol{X}) = 16$.

Table 2.2. Mean Deviations (*Observed* − *Population*) of Sphericity Estimates for the HF, Rank-Adjusted and GG with Population $\epsilon = 0.282$, Standard Error of Observed ≤ 0.0003.

N	Rank(X)	HF $\widetilde{\epsilon}$ (s.d.)	Rank-Adj $\widetilde{\epsilon}_r$ (s.d.)	GG $\widehat{\epsilon}$ (s.d.)
16	4	0.092 (0.033)	0.013 (0.027)	0.003 (0.020)
	8	0.314 (0.072)	0.022 (0.044)	0.005 (0.028)
32	4	0.037 (0.014)	0.005 (0.013)	0.001 (0.011)
	8	0.093 (0.018)	0.006 (0.014)	0.001 (0.012)
	16	0.296 (0.037)	0.009 (0.020)	0.002 (0.016)
48	4	0.023 (0.010)	0.003 (0.009)	0.001 (0.009)
	8	0.054 (0.011)	0.003 (0.010)	0.001 (0.009)
	16	0.142 (0.017)	0.004 (0.011)	0.001 (0.010)

In general, the same trends were also observed in Tables 2.3 and 2.4 for the population sphericity values of 0.505 and 0.720, respectively. For every case considered, the rank-adjusted sphericity estimates better estimated the population sphericity value than the Huynh-Feldt estimates, and the accuracy of the rank-adjusted estimates compared to those of the Huynh-Feldt seems to have improved as rank of $\boldsymbol{X}$ increased.

Table 2.3. Mean Deviations (*Observed* − *Population*) of Sphericity Estimates for the HF, Rank-Adjusted and GG with Population $\epsilon = 0.505$, Standard Error of Observed ≤ 0.0003.

N	Rank(X)	HF $\widetilde{\epsilon}$ (s.d.)	Rank-Adj $\widetilde{\epsilon}_r$ (s.d.)	GG $\widehat{\epsilon}$ (s.d.)
16	4	0.212 (0.160)	0.078 (0.150)	−0.019 (0.098)
	8	0.434 (0.096)	0.111 (0.185)	−0.034 (0.104)
32	4	0.091 (0.099)	0.034 (0.090)	−0.006 (0.074)
	8	0.196 (0.120)	0.039 (0.098)	−0.007 (0.079)
	16	0.446 (0.076)	0.059 (0.127)	−0.013 (0.091)
48	4	0.057 (0.073)	0.021 (0.068)	−0.003 (0.061)
	8	0.116 (0.085)	0.023 (0.072)	−0.004 (0.063)
	16	0.276 (0.109)	0.029 (0.083)	−0.005 (0.070)

Table 2.4. Mean Deviations (*Observed* − *Population*) of Sphericity Estimates for the HF, Rank-Adjusted and GG with Population $\epsilon = 0.720$, Standard Error of Observed ≤ 0.0003.

N	Rank(X)	HF $\widetilde{\epsilon}$ (s.d.)	Rank-Adj $\widetilde{\epsilon}_r$ (s.d.)	GG $\widehat{\epsilon}$ (s.d.)
16	4	0.186 (0.112)	0.045 (0.134)	−0.118 (0.086)
	8	0.272 (0.033)	0.059 (0.163)	−0.161 (0.093)
32	4	0.097 (0.088)	0.020 (0.081)	−0.057 (0.065)
	8	0.210 (0.080)	0.023 (0.089)	−0.066 (0.069)
	16	0.279 (0.010)	0.035 (0.114)	−0.093 (0.079)
48	4	0.062 (0.066)	0.012 (0.062)	−0.038 (0.053)
	8	0.140 (0.074)	0.013 (0.066)	−0.041 (0.056)
	16	0.265 (0.037)	0.017 (0.075)	−0.051 (0.061)

The Geisser-Greenhouse estimates are also presented in these tables in an attempt to demonstrate how well the new rank-adjusted estimator compares to one whose bias spawned the work of Huynh and Feldt (1976). In general, the Geisser-Greenhouse estimates seem to have better approximated the population sphericity values for the smaller values. However, as Huynh and Feldt (1976) had observed, the accuracy of the Geisser-Greenhouse estimates began to deteriorate as population sphericity increased. In the case of population sphericity of 0.720, the rank-adjusted sphericity estimates seem to have better approximated the population sphericity values than either of its two competitors.

Uniformly, both the Huynh-Feldt and new rank-adjusted sphericity estimators appear to have been biased high for population sphericity values of 0.282, 0.505 and 0.720. The Geisser-Greenhouse estimator seems to have been biased high when the population sphericity value was low, as evidenced by the results in Table 2.1 with population sphericity of 0.282. The bias became low when the population sphericity value increased, as evidenced by the results in Tables 2.2 and 2.3, with population sphericity values of 0.505 and 0.720, respectively.

In Table 2.5, the mean deviations between the Huynh-Feldt, rank-adjusted, and Geisser-Greenhouse sphericity estimates and the population sphericity value of 1.00 are presented for the various sample sizes and ranks of $\boldsymbol{X}$ considered. This table illustrates a case in which a researcher guesses incorrectly at the population sphericity and uses an approximate UNIREP test rather than the uncorrected test. In this case, the uncorrected tests would be uniformly most powerful among similarly invariant tests, and exact size alpha. In this case, obviously all sphericity estimators besides that of the uncorrected will be biased low. Of those examined here, the order of the sphericity estimators remain $\widehat{\epsilon} \leq \widetilde{\epsilon}_r \leq \widetilde{\epsilon}_{HF}$, which implies the Geisser-Greenhouse estimator will be most biased, followed by the rank-adjusted and the Huynh-Feldt. When the researcher does guess incorrectly about a spherical population, the rank-adjusted sphericity estimator still performs well. The largest deviation between the estimate and the population sphericity of 1.00 was -0.069 for those cases considered. The biases seem to have improved as sample size increased and rank of $\boldsymbol{X}$ decreased.

Table 2.5. Mean Deviations (*Observed* $-$ *Population*) of Sphericity Estimates for the HF, Rank-Adjusted and GG with Population $\epsilon = 1.00$, Standard Error of Observed ≤ 0.0003.

N	Rank(X)	HF $\widetilde{\epsilon}$ (s.d.)	Rank-Adj $\widetilde{\epsilon}_r$ (s.d.)	GG $\widehat{\epsilon}$ (s.d.)
16	4	$-0.007\,(0.030)$	$-0.051\,(0.082)$	$-0.255\,(0.081)$
	8	$-3.0 \times 10^{-4}\,(0.006)$	$-0.069\,(0.108)$	$-0.333\,(0.092)$
32	4	$-0.004\,(0.017)$	$-0.025\,(0.043)$	$-0.133\,(0.052)$
	8	$-2.0 \times 10^{-4}\,(0.004)$	$-0.029\,(0.049)$	$-0.151\,(0.057)$
	16	$-4.7 \times 10^{-7}\,(0.000)$	$-0.041\,(0.067)$	$-0.207\,(0.072)$
48	4	$-0.003\,(0.012)$	$-0.017\,(0.029)$	$-0.090\,(0.037)$
	8	$-2.0 \times 10^{-4}\,(0.003)$	$-0.019\,(0.032)$	$-0.098\,(0.040)$
	16	$-9.4 \times 10^{-7}\,(0.000)$	$-0.023\,(0.038)$	$-0.119\,(0.047)$

As described in section 2.3, both the Huynh-Feldt and rank-adjusted sphericity estimators are truncated when their estimates exceed a value of 1.0. The estimates are truncated to the maximum sphericity estimate value, 1.0. In Table 2.6, the proportions of the 500,000 observed Huynh-Feldt and rank-adjusted sphericity estimates that were truncated are shown for each sample size, rank of $\boldsymbol{X}$ and population sphericity value considered. As expected, virtually none of the estimates for the lowest population sphericity value considered were in need of truncation. As the population sphericity value increased, the need for truncation of both the Huynh-Feldt and rank-adjusted sphericity estimates increased as well. The Huynh-Feldt estimates were truncated much more frequently than the rank-adjusted estimates. The need for truncation of both estimators increased as sample size decreased and the rank of $\boldsymbol{X}$ increased. However, the proportion of truncations for the rank-adjusted estimator was much less affected by sample size and the rank of $\boldsymbol{X}$ than was the proportion of truncations for the Huynh-Feldt estimator.

Table 2.6. Proportions ($\times$ 100) of 500,000 Observed HF and Rank-Adjusted Sphericity Estimates Truncated to 1.0 for Sample Sizes, Rank(X) and Population Sphericities Considered.

		$\epsilon = 0.282$		$\epsilon = 0.505$		$\epsilon = 0.720$		$\epsilon = 1.00$	
N	Rank(X)	HF	RA	HF	RA	HF	RA	HF	RA
16	4	0.0	0.0	9.9	1.8	39.3	6.4	92.1	56.3
	8	0.5	0.0	60.7	6.4	92.7	15.5	99.6	56.4
32	4	0.0	0.0	0.1	0.0	2.3	0.1	91.5	56.2
	8	0.0	0.0	2.6	0.0	35.1	0.3	99.4	56.2
	16	0.0	0.0	59.2	0.5	98.4	2.4	100.0	56.3
48	4	0.0	0.0	0.0	0.0	0.1	0.0	91.3	56.2
	8	0.0	0.0	0.0	0.0	3.5	0.0	99.4	56.2
	16	0.0	0.0	5.5	0.0	77.3	0.0	100.0	56.2

As noted earlier, the Huynh-Feldt estimate will always be greater than or equal to the rank-adjusted estimate. This relationship corresponds to greater power values for the Huynh-Feldt test as compared to the rank-adjusted test, for a fixed $\boldsymbol{B}$ matrix of regression

coefficients. However, the greater power is a false power, purchased at the price of test size inflation. The uncorrected test will always have power values greater than or equal to the power values of the three corrected test. Yet, the uncorrected test is not used uniformly throughout UNIREP analyses because this increased power also comes with increased type I error rate. With target test size, α, of 0.05, Table 2.6 contains observed mean and predicted test sizes for a test of interaction for the Huynh-Feldt, the rank-adjusted and the Geisser-Greenhouse tests. Contrast matrices, $\boldsymbol{C}$ and $\boldsymbol{U}$, and $\boldsymbol{\Theta}_0$ are defined in Appendix C. Pseudo-random realizations of the error matrix, $\boldsymbol{E}$, were generated and appropriate test statistics were calculated. The observed mean test size for both the Huynh-Feldt and new rank-adjusted tests were calculated and tabulated under the null for $500{,}000$ replications per condition. The observed mean test size for each condition was the proportion of rejected tests among $500{,}000$ simulated realizations. All calculations were performed in SAS/IML. Observed test size values were computed using the modified version of LINMOD 3.3 for the null case. Predicted test size values were computed using a modified version of POWERLIB 2.0 (http://ehpr.ufl.edu/muller/), a software that computes statistical power for the General Linear Multivariate Model. POWERLIB 2.0 was modified to include the rank-adjusted test and will be made available soon.

In every case presented, with the exception of sphericity ($\epsilon = 1$), the observed mean and predicted rank-adjusted tests achieve a test size closer to the target test size than the corresponding Huynh-Feldt tests. In addition, the average error is merely slightly conservative for $\epsilon = 1$ for the observed mean rank-adjusted test. The largest observed mean and predicted Huynh-Feldt test sizes were 0.134 and 0.123, respectively. The maximum test sizes for the observed mean and predicted rank-adjusted test were only 0.079 and 0.050, respectively. In general, the larger test sizes for both the Huynh-Feldt and rank-adjusted tests were observed for mid-range population sphericity values. Test sizes seemed to increase with rank of $\boldsymbol{X}$ and decrease with sample size.

Table 2.7. Observed Mean and Predicted Interaction Test Size for Target $\alpha = 0.05$ for the HF, Rank-Adjusted and GG, Standard Error of Observed ≤ 0.0003. Degrees of freedom multipliers, $\tilde{\epsilon}_{HF}$, $\tilde{\epsilon}_r$ and $\hat{\epsilon}$, adjust for nonsphericity indexed by ϵ.

			Observed			Predicted		
N	Rank(X)	ϵ	HF	Rank-Adj	GG	HF	Rank-Adj	GG
16	4	0.282	0.075	0.056	0.053	0.068	0.050	0.048
		0.505	0.084	0.068	0.054	0.068	0.050	0.040
		0.720	0.071	0.059	0.040	0.068	0.050	0.034
		1.00	0.047	0.044	0.029	0.050	0.050	0.029
	8	0.282	0.134	0.064	0.056	0.123	0.050	0.047
		0.505	0.112	0.079	0.051	0.117	0.050	0.034
		0.720	0.082	0.065	0.036	0.082	0.050	0.027
		1.00	0.052	0.046	0.024	0.050	0.050	0.021
32	4	0.282	0.062	0.055	0.054	0.057	0.050	0.049
		0.505	0.064	0.057	0.052	0.057	0.050	0.045
		0.720	0.062	0.055	0.047	0.058	0.050	0.043
		1.00	0.049	0.047	0.039	0.050	0.050	0.040
	8	0.282	0.076	0.054	0.053	0.072	0.050	0.049
		0.505	0.085	0.064	0.056	0.072	0.050	0.044
		0.720	0.075	0.057	0.046	0.072	0.050	0.041
		1.00	0.050	0.048	0.037	0.050	0.050	0.037
	16	0.282	0.128	0.057	0.054	0.123	0.050	0.048
		0.505	0.118	0.071	0.057	0.120	0.050	0.041
		0.720	0.082	0.060	0.043	0.082	0.050	0.036
		1.00	0.050	0.046	0.032	0.050	0.050	0.032
48	4	0.282	0.056	0.051	0.050	0.054	0.050	0.050
		0.505	0.063	0.058	0.055	0.055	0.050	0.047
		0.720	0.059	0.054	0.049	0.055	0.050	0.045
		1.00	0.047	0.046	0.041	0.050	0.050	0.043
	8	0.282	0.068	0.055	0.055	0.063	0.050	0.049
		0.505	0.072	0.060	0.056	0.063	0.050	0.046
		0.720	0.068	0.056	0.047	0.063	0.050	0.044
		1.00	0.049	0.047	0.047	0.050	0.050	0.042
	16	0.282	0.087	0.053	0.052	0.086	0.050	0.049
		0.505	0.098	0.063	0.057	0.086	0.050	0.045
		0.720	0.082	0.057	0.049	0.080	0.050	0.043
		1.00	0.048	0.046	0.038	0.050	0.050	0.040

2.5 Conclusions

In UNIREP analyses, statisticians are rarely able to claim sphericity with confidence unless much is known about the data beforehand. As a result, approximate UNIREP tests, such as the Geisser-Greenhouse and the Huynh-Feldt, are often employed. To use these tests, appropriate sphericity estimates are needed. The Huynh-Feldt sphericity estimator was developed in an attempt to correct for biases found with the Geisser-Greenhouse estimator. Huynh and Feldt (1976) claimed that their estimator was a ratio of unbiased estimators. They further asserted that their estimator was less biased and less dependent on large sample sizes than the Geisser-Greenhouse estimator when the population covariance deviated only moderately from sphericity. As demonstrated in this paper, their claims are only true for the special case of rank of $\boldsymbol{X}$ equal to 1. The rank-adjusted estimator, however, is a ratio of unbiased estimators for any rank of $\boldsymbol{X}$, reducing to that of the Huynh-Feldt when rank of $\boldsymbol{X}$ is equal to 1. In all situations in which use of an approximate UNIREP test would be called for, and for all sample sizes and all rank of $\boldsymbol{X}$ greater than 1 considered, the rank-adjusted sphericity estimator better estimated the population sphericity than the Huynh-Feldt. This outcome was particularly true for larger rank of $\boldsymbol{X}$.

Although both the rank-adjusted and Huynh-Feldt estimators seem to be biased high when the population is not completely spherical, this bias for the rank-adjusted estimator seems to be less than that of the Geisser-Greenhouse estimator as the population approaches sphericity. This trend is particularly obvious in the case of a spherical population, in which a researcher guesses incorrectly at the population sphericity and uses an approximate UNIREP test instead of using the uncorrected test. The uncorrected test would be uniformly most powerful among similarly invariant tests, and exact size alpha in the case of a spherical population.

The Huynh-Feldt estimate will always be greater than or equal to that of the rank-adjusted estimate. This relationship results in greater power for the Huynh-Feldt test when

compared to the rank-adjusted for a fixed $\boldsymbol{B}$ matrix of regression coefficients. However, this increased power comes at a price. The Huynh-Feldt test demonstrated inflated test size for many of the cases considered in this paper. In some cases, the simulated Huynh-Feldt test reached a test size greater than double the target size. The rank-adjusted test essentially always controlled test size adequately.

Based on the results presented here, the new rank-adjusted sphericity estimator is recommended to immediately replace the Huynh-Feldt estimator in all statistical software that handles UNIREP analyses. In all cases except for a spherical population, when an approximate UNIREP test would not be called for, the rank-adjusted estimator far outperformed that of the Huynh-Feldt. Furthermore, its use is more theoretically in line with the goals originally set forth by Huynh and Feldt (1976).

Chapter 3
Approximate Power for a More General Version of the Huynh-Feldt Test

3.1 Motivation

When sphericity is not met, Box (1954a, b) showed that the UNIREP test statistic under the null could be approximately distributed as an F with reduced degrees of freedom, $F(ab\epsilon, ab\nu_e\epsilon)$. The reduction comes from various estimators of ϵ, a measure of sphericity, used as degrees of freedom multipliers. Geisser and Greenhouse (1958) offered the Box conservative estimate, $\epsilon = 1/b$, and the maximum likelihood estimator, $\widehat{\epsilon} = [\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)]/[b\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)]$, now known as the Geisser-Greenhouse estimator, as degrees of freedom multipliers.

In their 1976 paper, Huynh and Feldt described examples in which the Geisser-Greenhouse estimator was seriously biased, most often in nearly spherical populations. They found that in such cases the estimator overcorrected the degrees of freedom, resulting in a more stringent significance level than the nominal level desired. Huynh and Feldt (1976) responded with an estimator of ϵ, which they showed through examples to be less biased and less dependent on large sample sizes than the Geisser-Greenhouse estimator when the population covariance deviated only moderately from sphericity. They claimed that their estimator, $\widetilde{\epsilon} = (Nb\widehat{\epsilon} - 2)/[b(\nu_e - b\widehat{\epsilon})]$, was a ratio of unbiased estimators. In Chapter 2, this claim was demonstrated to be false, except for the special case of rank of $\boldsymbol{X}$ equal to 1. A new rank-adjusted approximately unbiased estimator, $\widetilde{\epsilon}_r$, was proposed and shown to be a better estimator of the population sphericity parameter than the Huynh-Feldt

estimator when rank of $\boldsymbol{X}$ is greater than 1, and equal to the Huynh-Feldt estimator when rank of $\boldsymbol{X}$ is equal to 1.

A power approximation for the rank-adjusted test using the rank-adjusted sphericity estimator will be proposed. For practical research situations, this power approximation will be shown to be as accurate as the Huynh-Feldt power approximation introduced by Muller *et al.* (2007), which incorporates the Huynh-Feldt sphericity estimator. Furthermore, the rank-adjusted approximation will be shown to adequately control test size when rank of $\boldsymbol{X}$ is greater than 1. In comparison, the Huynh-Feldt approximation will be shown to result in artificially inflated power and inflated test size when rank of $\boldsymbol{X}$ is greater than 1, thus potentially biasing the study analysis. The severity of the artificial inflation will be examined. The Huynh-Feldt estimator, $\tilde{\epsilon}$, will hence forth be referred to as $\tilde{\epsilon}_{HF}$.

3.2 Notation and Known Results

Muller *et al.* (2007) introduced power approximations for all four UNIREP tests. Using simulations, they demonstrated that their power approximations eliminated most inaccuracy present in existing methods. Muller *et al.* (2007, Appendix A) showed that the UNIREP test statistic CDF could be expressed exactly for the Box conservative and the uncorrected tests. Due to their random critical values, the CDFs for the Geisser-Greenhouse and Huynh-Feldt test statistics could only be approximated. Muller *et al.* (2007) formulated approximations for the expected values of both the Geisser-Greenhouse and Huynh-Feldt sphericity estimators for use in power calculations. Only their approximation of the expected value of the Huynh-Feldt sphericity estimator, $\mathrm{E}(\tilde{\epsilon}_{HF})$, for the Huynh-Feldt test is pertinent to this discussion, and only the power approximations presented in Muller *et al.* (2007) will be considered here.

Their power approximation for the Huynh-Feldt test uses an approximate value of $\mathrm{E}[t_{\text{crit}}(\text{HF})]$ as the critical value, such that $\mathrm{E}[t_{\text{crit}}(\text{HF})] \approx$ $F_F^{-1}[1-\alpha; ab\mathrm{E}(\tilde{\epsilon}_{HF}), b\nu_e\mathrm{E}(\tilde{\epsilon}_{HF})]$. Here, a and b are the ranks of the between- and

within-subject contrast matrices, $\boldsymbol{C}$ and $\boldsymbol{U}$, respectively, and ν_e is the error degrees of freedom between subjects, $\nu_e = N - \text{rank}(\boldsymbol{X})$. Muller *et al.* (2007) expressed the Huynh-Feldt estimator as $\widetilde{\epsilon}_{HF} = b^{-1}(Nt_1 - 2t_2)/(\nu_e t_2 - t_1)$, such that $t_1 = \text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)$ and $t_2 = \text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)$. They showed that if $d = (\nu_e t_2 - t_1)$, then a first order Taylor series for d^{-1} about the point d_0 gives

$$\text{E}(\widetilde{\epsilon}_{HF}) \approx b^{-1}\text{E}\big\{(Nt_1 - 2t_2)\big[d_0^{-1} - d_0^{-2}(d - d_0)\big]\big\} \,. \tag{24}$$

With $\overline{\lambda}$ equal to $\text{tr}(\boldsymbol{\Sigma}_*)/b$, the Huynh-Feldt power approximation is of the form

$$P = 1 - F_F\left[F_F^{-1}(1-\alpha; \text{E}(\widetilde{\epsilon}_{HF}) \cdot ab, \text{E}(\widetilde{\epsilon}_{HF}) \cdot b\nu_e); \epsilon_n \cdot ab, \epsilon_d \cdot \nu_e b, \frac{\text{tr}(\boldsymbol{\Delta})}{\overline{\lambda}/\epsilon_n}\right], \tag{25}$$

such that

$$\epsilon_n = \frac{\text{tr}^2(\boldsymbol{\Sigma}_*) + 2\text{tr}(\boldsymbol{\Sigma}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b[\text{tr}(\boldsymbol{\Sigma}_*^2) + 2\text{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta}/a)]} \tag{26}$$

$$\epsilon_d = \frac{\text{tr}^2(\boldsymbol{\Sigma}_*)}{b\text{tr}(\boldsymbol{\Sigma}_*^2)} = \epsilon_n|\boldsymbol{\Delta} = \boldsymbol{0} \qquad \equiv \epsilon \,. \tag{27}$$

The parameter ϵ_n is the sphericity parameter under the nonnull case. As depicted above in equation 27, ϵ_n reduces to the familiar sphericity parameter under the null case.

In section 2.3, a rank-adjusted approximately unbiased estimator was introduced and shown to better estimate the population sphericity parameter than the Huynh-Feldt estimator in all cases except sphericity ($\epsilon = 1$). In the spherical case, neither the Huynh-Feldt nor the rank-adjusted tests would be considered. The rank-adjusted estimator achieves what Huynh and Feldt (1976) had originally intended to accomplish. The rank-adjusted estimator is a ratio of two unbiased estimators, which is less biased and less dependent on large sample sizes than the Geisser-Greenhouse estimator when the covariance matrix deviates only moderately from sphericity, and is so for any rank of $\boldsymbol{X}$. The rank-adjusted estimator is

$$\tilde{\epsilon}_r = \frac{(\nu_e + 1)\hat{b\epsilon} - 2}{b(\nu_e - \hat{b\epsilon})} \ . \tag{28}$$

Like the Huynh-Feldt estimator, the rank-adjusted estimator is truncated to 1.0 when necessary. The rank-adjusted estimator will always be less than or equal to the Huynh-Feldt estimator, with equality only if rank of $\boldsymbol{X}$ is equal to 1. This relationship corresponds to greater Huynh-Feldt power values than rank-adjusted power values, for a fixed $\boldsymbol{B}$ matrix of regression coefficients. However, due to the poor ability of the Huynh-Feldt estimator to accurately estimate the population sphericity as compared to the rank-adjusted estimator when rank of $\boldsymbol{X}$ is greater than 1, these power values are artificially inflated, and may lead to bias in a study analysis.

3.3 Power Approximation for a More General Version of the Huynh-Feldt Test

Using the same notation introduced in Muller *et al.* (2007), the rank-adjusted sphericity estimator can be expressed as $\tilde{\epsilon}_r = b^{-1}[(\nu_e + 1)t_1 - 2t_2]/(\nu_e t_2 - t_1)$. The expected value of the rank-adjusted estimator can be approximated with a first order Taylor series for d^{-1} about the point d_0 as

$$\mathrm{E}(\tilde{\epsilon}_r) \approx b^{-1}\mathrm{E}\big\{[(\nu_e + 1)t_1 - 2t_2]\big[d_0^{-1} - d_0^{-2}(d - d_0)\big]\big\} \ , \tag{29}$$

such that $d = (\nu_e t_2 - t_1)$. The approximation method uses one term in equation 29 and $d_0 = \mathrm{E}(d)$ to give

$$\mathrm{E}(\tilde{\epsilon}_r) \approx \frac{(\nu_e + 1)\mathrm{E}(t_1) - 2\mathrm{E}(t_2)}{b[\nu_e \mathrm{E}(t_2) - \mathrm{E}(t_1)]} \ . \tag{30}$$

In section 2.2, the expected values of t_1 and t_2 are given to be

$$\mathrm{E}(t_1) = 2\nu_e \mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e^2 \mathrm{tr}^2(\boldsymbol{\Sigma}_*) \tag{31}$$

and

$$\mathrm{E}(t_2) = \nu_e(\nu_e + 1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e \mathrm{tr}^2(\boldsymbol{\Sigma}_*) \ . \tag{32}$$

Thus, the approximate expected value of the rank-adjusted sphericity estimator is

$$\mathrm{E}(\widetilde{\epsilon}_r) \approx \frac{(\nu_e+1)[2\nu_e \mathrm{tr}(\boldsymbol{\Sigma}_*^2)+\nu_e^2 \mathrm{tr}^2(\boldsymbol{\Sigma}_*)]-2[\nu_e(\nu_e+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2)+\nu_e \mathrm{tr}^2(\boldsymbol{\Sigma}_*)]}{b\{\nu_e[\nu_e(\nu_e+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2)+\nu_e \mathrm{tr}^2(\boldsymbol{\Sigma}_*)]-[2\nu_e \mathrm{tr}(\boldsymbol{\Sigma}_*^2)+\nu_e^2 \mathrm{tr}^2(\boldsymbol{\Sigma}_*)]\}} \,. \quad (33)$$

This expected value approximation can be incorporated into the power approximation as presented in equation 25, such that

$$P = 1 - F_F\left[F_F^{-1}(1-\alpha; \mathrm{E}(\widetilde{\epsilon}_r)\cdot ab, \mathrm{E}(\widetilde{\epsilon}_r)\cdot b\nu_e); \epsilon_n \cdot ab, \epsilon_d \cdot \nu_e b, \frac{\mathrm{tr}(\boldsymbol{\Delta})}{\overline{\lambda}/\epsilon_n}\right] . \quad (34)$$

When the rank of $\boldsymbol{X}$ equals 1, the rank-adjusted sphericity estimator reduces to the Huynh-Feldt sphericity estimator. The reduction guarantees exactly the same tests and approximate power calculations for the special case.

3.4 Simulations

The power approximation using the rank-adjusted approximately unbiased sphericity estimator was evaluated for a wide range set of simulations. All of the simulations considered the condition of rank of $\boldsymbol{X}$ greater than 1. When rank of $\boldsymbol{X}$ is equal to 1, the Huynh-Feldt and rank-adjusted power approximations are identical. The simulated realizations consisted of $p = 5$ repeated measures, N the sample size of 16, 32 and 48, and q the rank of $\boldsymbol{X}$ equal to 4, 8 and 16, in the model

$$\underset{(N\times 5)}{\boldsymbol{Y}} = \underset{(N\times q\times 5)}{\boldsymbol{XB}} + \underset{(N\times 5)}{\boldsymbol{E}} \,. \quad (35)$$

Appropriate fixed matrices of regression coefficients, $\boldsymbol{B}$, and contrast matrices, $\boldsymbol{C}$ and $\boldsymbol{U}$, and $\boldsymbol{\Theta}_0$ were chosen to test an interaction for a test size, α, of 0.05, and to ensure target predicted power values for the rank-adjusted test of 0.20, 0.50 and 0.80. Specific design matrices, $\boldsymbol{X}$, were defined. Population covariance matrices were chosen to provide specific population sphericity values, $\epsilon \in \{0.282, 0.505, 0.720, 1.00\}$.

Pseudo-random realizations of the error matrix, $\boldsymbol{E}$, were generated and appropriate test statistics were calculated. The observed mean power values for both the Huynh-Feldt and rank-adjusted tests were calculated and tabulated for 500,000 replications per condition.

The observed mean power values for each condition was the proportion of rejected tests among 500,000 simulated realizations. Observed power values were computed using a modified version of LINMOD 3.3 (http://ehpr.ufl.edu/muller/), a software that performs a wide variety of General Linear Multivariate Model computations. LINMOD 3.3 was modified to include the rank-adjusted estimator and test.

The rank-adjusted predicted power values were calculated using the approximation introduced in section 3.3 above. Predicted power values for the Huynh-Feldt test were calculated as well using the approximation defined in section 3.2. Predicted power values were computed using a modified version of POWERLIB 2.0 (http://ehpr.ufl.edu/muller/), a software that computes statistical power for the General Linear Multivariate Model. POWERLIB 2.0 was modified to include the rank-adjusted test. The modified versions of both LINMOD and POWERLIB will be made available soon.

In Table 3.1, the mean power deviations of the predicted and observed mean rank-adjusted and Huynh-Feldt tests, respectively, are tabulated for the rank of $\boldsymbol{X}$ and target power values considered for sample size of 16 and for a population sphericity of 0.282. The power values and deviations have been multiplied by 100 for clarity of presentation. Similar results for population sphericity values of 0.505, 0.720 and 1.00 are presented in Tables 3.2-3.4.

In every case considered, the mean deviations between the predicted and observed mean rank-adjusted and predicted and observed mean Huynh-Feldt power values were adequately small. For population sphericity of 0.282, the largest absolute observed mean deviation observed for the rank-adjusted and Huynh-Feldt tests were 0.027 and 0.020, respectively. For population sphericity of 0.505, the largest absolute observed mean deviation observed for the rank-adjusted and Huynh-Feldt tests were 0.064 and 0.023, respectively. For population sphericity of 0.720, the largest absolute observed mean deviation observed for the rank-adjusted and Huynh-Feldt tests were 0.023 and 0.005,

respectively. The most severe deviations of the observed mean and predicted power values for the rank-adjusted test occurred for the smallest target power value, 0.20.

Table 3.1. Predicted and Observed Mean Rank-Adjusted and HF Power ($\times$ 100) for Rank$(X) = q$ and $N = 16$, Standard Error of Observed < 0.001. Mean Differences of Predicted and Observed Rank-Adjusted and HF Power ($\times$ 100) for Population $\epsilon = 0.282$.

	Predicted Power		Observed Power (Mean)		Observed − Predicted Mean	
q	Rank-Adj.	HF	Rank-Adj.	HF	Rank-Adj.	HF
2	20	21.6	18.2	19.6	-1.8	-2.0
	50	52.6	49.5	52.3	-0.5	-0.3
	80	82.1	81.7	83.9	1.7	1.8
4	20	26.0	20.7	26.2	0.7	0.2
	50	59.1	50.3	59.3	0.3	0.2
	80	86.5	80.5	87.1	0.5	0.6
8	20	39.7	22.7	41.0	2.7	1.3
	50	75.1	51.7	75.4	1.7	0.3
	80	94.8	79.9	94.7	-0.1	-0.1

Table 3.2. Predicted and Observed Mean Rank-Adjusted and HF Power ($\times$ 100) for Rank$(X) = q$ and $N = 16$, Standard Error of Observed < 0.001. Mean Differences of Predicted and Observed Rank-Adjusted and HF Power ($\times$ 100) for Population $\epsilon = 0.505$.

	Predicted Power		Observed Power (Mean)		Observed − Predicted Mean	
q	Rank-Adj.	HF	Rank-Adj.	HF	Rank-Adj.	HF
2	20	21.4	21.3	22.6	1.3	1.2
	50	52.2	51.9	54.0	1.9	1.8
	80	81.7	80.9	82.5	0.9	0.8
4	20	25.3	23.1	27.6	3.1	2.3
	50	57.7	52.7	59.4	2.7	1.7
	80	85.3	80.2	85.3	0.2	0.0
8	20	36.9	26.4	36.8	6.4	-0.1
	50	71.3	55.0	70.6	5.0	-0.7
	80	92.8	79.6	91.5	-0.4	-1.3

Table 3.3. Predicted and Observed Mean Rank-Adjusted and HF Power ($\times$ 100) for Rank$(X) = q$ and $N = 16$, Standard Error of Observed < 0.001. Mean Differences of Predicted and Observed Rank-Adjusted and HF Power ($\times$ 100) for Population $\epsilon = 0.720$.

	Predicted Power		Observed Power (Mean)		Observed − Predicted Mean	
q	Rank-Adj.	HF	Rank-Adj.	HF	Rank-Adj.	HF
2	20	21.3	20.2	21.3	0.2	0.0
	50	51.9	50.1	51.9	0.1	0.0
	80	81.3	80.5	81.8	0.5	0.5
4	20	24.9	21.2	24.8	1.2	-0.1
	50	56.9	50.8	56.4	0.8	-0.5
	80	84.7	80.1	84.4	0.1	-0.3
8	20	28.3	22.3	28.0	2.3	-0.3
	50	61.3	51.9	61.2	1.9	-0.1
	80	87.5	79.5	87.4	-0.5	-0.1

Table 3.4. Predicted and Observed Mean Rank-Adjusted and HF Power ($\times$ 100) for Rank$(X) = q$ and $N = 16$, Standard Error of Observed < 0.001. Mean Differences of Predicted and Observed Rank-Adjusted and HF Power ($\times$ 100) for Population $\epsilon = 1.00$.

	Predicted Power		Observed Power (Mean)		Observed − Predicted Mean	
q	Rank-Adj.	HF	Rank-Adj.	HF	Rank-Adj.	HF
2	20	20	19.2	19.6	-0.8	-0.4
	50	50	48.7	49.3	-1.3	-0.7
	80	80	79.0	79.5	-1.0	0.5
4	20	20	18.8	19.8	-1.2	-0.2
	50	50	48.3	49.9	-1.7	-0.1
	80	80	78.7	79.9	-1.3	-0.1
8	20	20	18.2	20.0	-1.8	0.0
	50	50	47.0	50.0	-3.0	0.0
	80	80	77.5	79.9	-2.5	-0.1

In practical biomedical research, lower power values are of little concern. Rarely will one have a power analyses targeted below 0.70. Thus, an examination of how well the rank-adjusted and Huynh-Feldt power approximations perform at the target power of 0.80

would be beneficial. Additional predicted and observed mean power values for the rank-adjusted and Huynh-Feldt tests for target power of 0.80 are tabulated in Table 3.5. The cases considered consisted of population sphericity values of 0.282, 0.505 and 0.720, sample sizes of 32 and 48 and rank of $\boldsymbol{X}$ greater than 1. In nearly every case considered, the new rank-adjusted predicted power approximation better approximated the associated observed mean power values than did the Huynh-Feldt predicted power approximation. The power approximations for both tests performed extremely well in all cases. The largest absolute mean deviations observed for the rank-adjusted and Huynh-Feldt tests were only 0.045 and 0.046, respectively. The deviations for both tests decreased as rank of $\boldsymbol{X}$ and population sphericity increased and as sample size decreased.

Table 3.4, with population sphericity of 1.00, is presented solely for the sake of being complete with respect to the research at hand. In practice, if one believes that the population is spherical, neither the rank-adjusted nor the Huynh-Feldt tests would be used. This table illustrates a case in which a researcher guesses incorrectly at the population sphericity and, instead of using the uncorrected test which would be uniformly most powerful among similarly invariant tests, and exact size alpha, uses an approximate UNIREP test. A higher incidence of truncation allows the HF test to achieve a higher power, which is acceptable here because $\epsilon = 1$. However, the uncorrected test would be appropriately even better.

Table 3.5. Predicted and Observed Mean Rank-Adjusted and HF Power ($\times$ 100) for Rank$(X) = q$ and $N \in (32, 48)$, Standard Error of Observed < 0.001. Mean Differences of Predicted and Observed Rank-Adjusted and HF Power ($\times$ 100).

			Predicted Power		Observed Power (Mean)		Observed − Predicted Mean	
ϵ	N	q	Rank-Adj.	HF	Rank-Adj.	HF	Rank-Adj.	HF
0.282	32	2	80	80.9	83.5	84.5	3.5	3.6
		4	80	82.8	81.1	84.3	1.1	1.5
		8	80	86.5	80.2	86.9	0.2	0.4
		16	80	93.6	79.9	93.6	-0.1	0.0
	48	2	80	80.6	84.5	85.2	4.5	4.6
		4	80	81.8	81.5	83.6	1.5	1.8
		8	80	84.1	80.3	84.7	0.3	0.6
		16	80	88.8	79.9	88.9	-0.1	0.1
0.505	32	2	80	80.7	81.2	81.9	1.2	1.2
		4	80	82.3	80.4	82.8	0.4	0.5
		8	80	85.6	79.8	85.4	-0.2	-0.2
		16	80	92.2	79.4	91.2	-0.6	-1.0
	48	2	80	80.5	81.3	81.8	1.3	1.3
		4	80	81.5	80.5	82.1	0.5	0.6
		8	80	83.5	79.9	83.6	-0.1	0.1
		16	80	87.8	79.5	87.4	-0.5	-0.4
0.720	32	2	80	80.6	80.5	81.2	0.5	0.6
		4	80	82.1	80.2	82.3	0.2	0.2
		8	80	85.1	79.7	84.7	-0.3	-0.4
		16	80	87.0	79.6	87.1	-0.4	0.1
	48	2	80	80.4	80.6	81.0	0.6	0.6
		4	80	81.3	80.2	81.7	0.2	0.4
		8	80	83.2	79.9	83.2	-0.1	0.0
		16	80	86.5	79.9	86.3	-0.1	-0.2

The Huynh-Feldt sphericity estimate will always be greater than or equal to the rank-adjusted sphericity estimate, as demonstrated through their respective equations and simulations in Chapter 2. In turn, this relationship corresponds to Huynh-Feldt power

values greater than or equal to equivalent rank-adjusted power values for a fixed $\boldsymbol{B}$ matrix of regression coefficients. For example, in Table 3.5, for population sphericity of 0.282, sample size of 32, rank of $\boldsymbol{X}$ equal to 16 and a rank-adjusted predicted power of 0.80, the Huynh-Feldt predicted power was 0.936. The difference in predicted powers is 0.136. For similar conditions for population sphericity values of 0.505 and 0.720, the Huynh-Feldt predicted powers were 0.912 and 0.871. This increased power seems to have decreased as population sphericity and sample size increased and rank of $\boldsymbol{X}$ decreased. These power values are artificially inflated, however. The uncorrected test will always have power values greater than or equal to the power values of the three corrected test. Yet, the uncorrected test is not used uniformly throughout UNIREP analyses because this increased power also comes with an increased type I error rate.

With target test size, α, of 0.05, Table 3.6 contains observed mean and predicted test sizes for a test of interaction for the Huynh-Feldt and the rank-adjusted tests. Contrast matrices, $\boldsymbol{C}$ and $\boldsymbol{U}$, and $\boldsymbol{\Theta}_0$ are defined in Appendix C. Pseudo-random realizations of the error matrix, $\boldsymbol{E}$, were generated and appropriate test statistics were calculated. The observed mean test size for both the Huynh-Feldt and new rank-adjusted tests were calculated and tabulated under the null for 500,000 replications per condition. The observed mean test size for each condition was the proportion of rejected tests among 500,000 simulated realizations. All calculations were performed in SAS/IML. Observed test size values were computed using the modified version of LINMOD 3.3 for the null case. Predicted test size values were computed using the modified version of POWERLIB 2.0 under the same assumption.

In every case presented, with the exception of sphericity ($\epsilon = 1$), the observed mean and predicted rank-adjusted tests achieve a test size closer to the target than the observed mean and predicted Huynh-Feldt, respectively. The largest observed mean and predicted Huynh-Feldt test sizes among the conditions considered were 0.134 and 0.123, respectively,

while the largest observed mean rank-adjusted test size was only 0.079. In general, the larger test sizes for both the Huynh-Feldt and rank-adjusted tests were observed for mid-range population sphericity values. The test sizes seemed to increase with rank of $\boldsymbol{X}$ and decrease with sample size. The severity of inflated test size seemed to increase for the Huynh-Feldt test with rank of $\boldsymbol{X}$. The rank-adjusted test maintained control over test size as rank of $\boldsymbol{X}$ increased.

Table 3.6. Observed Mean and Predicted Interaction Test Size for Target $\alpha = 0.05$ for Rank$(X) = q$ for the HF and Rank-Adjusted, Standard Error of Observed ≤ 0.0003. Degrees of freedom multipliers, $\tilde{\epsilon}_{HF}$ and $\tilde{\epsilon}_r$, adjust for nonsphericity indexed by ϵ.

			Observed		Predicted	
N	q	ϵ	HF	Rank-Adj	HF	Rank-Adj
16	4	0.282	0.075	0.056	0.068	0.050
		0.505	0.084	0.068	0.068	0.050
		0.720	0.071	0.059	0.068	0.050
		1.00	0.047	0.044	0.050	0.050
	8	0.282	0.134	0.064	0.123	0.050
		0.505	0.112	0.079	0.117	0.050
		0.720	0.082	0.065	0.082	0.050
		1.00	0.052	0.046	0.050	0.050
32	4	0.282	0.062	0.055	0.057	0.050
		0.505	0.064	0.057	0.057	0.050
		0.720	0.062	0.055	0.058	0.050
		1.00	0.049	0.047	0.050	0.050
	8	0.282	0.076	0.054	0.072	0.050
		0.505	0.085	0.064	0.072	0.050
		0.720	0.075	0.057	0.072	0.050
		1.00	0.050	0.048	0.050	0.050
	16	0.282	0.128	0.057	0.123	0.050
		0.505	0.118	0.071	0.120	0.050
		0.720	0.082	0.060	0.082	0.050
		1.00	0.050	0.046	0.050	0.050
48	4	0.282	0.056	0.051	0.054	0.050
		0.505	0.063	0.058	0.055	0.050
		0.720	0.059	0.054	0.055	0.050
		1.00	0.047	0.046	0.050	0.050
	8	0.282	0.068	0.055	0.063	0.050
		0.505	0.072	0.060	0.063	0.050
		0.720	0.068	0.056	0.063	0.050
		1.00	0.049	0.047	0.050	0.050
	16	0.282	0.087	0.053	0.086	0.050
		0.505	0.098	0.063	0.086	0.050
		0.720	0.082	0.057	0.080	0.050
		1.00	0.048	0.046	0.050	0.050

3.5 Conclusions

In section 2.3, a rank-adjusted approximately unbiased sphericity estimator was proposed. The rank-adjusted estimator achieved what Huynh and Feldt (1976) had originally intended to accomplish. The rank-adjusted estimator is a ratio of two unbiased estimators, which is less biased and less dependent on large sample sizes than the Geisser-Greenhouse estimator when the covariance matrix deviates only moderately from sphericity, and is so for any rank of $\boldsymbol{X}$. In section 2.4 through simulations, the rank-adjusted estimator is shown to better estimate the population sphericity than the Huynh-Feldt estimator when rank of $\boldsymbol{X}$ is greater than 1. The rank-adjusted estimator reduces to the Huynh-Feldt estimator when the rank of $\boldsymbol{X}$ equals 1.

Muller *et al.* (2007) incorporated the Huynh-Feldt estimator into their power approximation for the Huynh-Feldt test. The, in some cases severe, bias that accompanies the Huynh-Feldt estimator yields artificially inflated power values for the Huynh-Feldt test. The artificially inflated power values for the Huynh-Feldt test could greatly bias any study analysis that utilizes the Huynh-Feldt test. The newly introduced rank-adjusted power approximation is more in line with the original intent of the research of Huynh and Feldt (1976), and more accurately depicts the true power of the test. The rank-adjusted power approximation deviated only minimally from the observed mean test powers for the cases considered. The most severe deviations of the predicted power using the rank-adjusted power approximation from the observed mean power values occurred for small target power values. In practice, smaller power values are irrelevant. In biomedical research, target powers smaller than 0.70 are rarely of interest. For the larger target power values, the rank-adjusted power approximation matched the observed mean power values as well as, or better than, the corresponding Huynh-Feldt approximations. In the case of a spherical population, the rank-adjusted and Huynh-Feldt power approximations were identical in all cases, most likely due to the truncation of the sphericity estimators.

Preferring the Huynh-Feldt power approximation simply because it provides a greater power when compared to the rank-adjusted for a fixed $\boldsymbol{B}$ matrix of regression coefficients would be a mistake. The increased power comes at a price. The Huynh-Feldt test of interaction demonstrated inflated test size over the rank-adjusted test in every case considered. In some cases, the Huynh-Feldt test size was greater than double the target size. The rank-adjusted test controlled test size adequately.

Based on the results presented here, use of the rank-adjusted power approximation along with the rank-adjusted estimator and test are recommended over the corresponding Huynh-Feldt methods. In all cases when an approximate UNIREP test would be called for, the rank-adjusted power approximations accurately described the true nature of the test. In fact, the rank-adjusted power approximations performed as well as and, in most cases, better than the Huynh-Feldt power approximations for practical target power values. They also controlled test size better than the Huynh-Feldt in all cases considered. Furthermore, use of the rank-adjusted test is more theoretically in line with the original intent of the research of Huynh and Feldt (1976).

Chapter 4
Power Confidence Intervals for UNIREP Tests

4.1 Motivation

Imaging is used in all areas of medical research. The number of medical applications seems to increase every year, while the cost of such procedures decreases. Researchers and physicians alike are realizing the benefits of using these safe and non-invasive techniques.

Imaging research often generates the type of complete data that can be handled with UNIREP procedures. UNIREP makes up a special case of the more broad area of statistical modeling called mixed models. The mixed model has several nice statistical features, such as no requirement for balanced data, the ability to explicitly model and analyze the between- and within-subject variation, and the capability of handling missing data without eliminating all values for a particular subject. However, there is still a need for better inference and power analysis techniques in mixed models. This is much less true for UNIREP. The inference techniques for UNIREP far outshine those used in mixed models, particularly for small sample sizes, and power techniques for UNIREP have been well tested and documented.

The power of a test is the probability of rejecting the null hypothesis. More and more, researchers are realizing the need for accurate power analysis and the role it plays in focusing the hypothesis, clarifying the analysis plan and enhancing study design efficiency. Power is computed assuming known values of distributional parameters. Rarely is the variance actually known in these computations. Rather, the variance is often estimated from previous studies. The estimated variance leads to random power values for a fixed sample size. Providing confidence intervals to account for the uncertainty inherent in the

random power values would be useful in any study design. A lower bound for power would allow stating that a study has power of at least "P" to detect an effect, with a specified confidence.

Hypothesis tests following the univariate approach to repeated measures are called UNIREP tests. The UNIREP tests include four types: the Box conservative, the Geisser-Greenhouse, the Huynh-Feldt, and the uncorrected. For data analysis, UNIREP tests differ only by their respective degrees of freedom due to different degrees of freedom multipliers, which are measures of the sphericity in the covariance model.

Taylor and Muller (1995) demonstrated how to construct exact power confidence intervals for the general linear univariate model for an estimated variance and fixed means. Despite the prevalence of such designs in practice, methods to provide accurate confidence intervals for power of a test in a UNIREP setting do not exist. In this paper, the methods introduced by Taylor and Muller (1995) are built upon and applied to UNIREP tests. Furthermore, the techniques proposed are shown to allow for the calculation of accurate, approximate confidence intervals for the UNIREP tests, in the case of an estimated covariance and fixed means.

The methods presented here focus solely on UNIREP procedures. Ultimately, one would hope to be able to apply these methods to all forms of mixed models. The expectation is that the research presented here will lay the groundwork for future research that will extend these methods to fit with the general mixed model.

4.2 Notation and Known Results: UNIREP Power Approximations

The univariate approach to repeated measures can be expressed in terms of the General Linear Multivariate Model (GLMM),

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}\,, \tag{36}$$

such that $\boldsymbol{Y}$ and $\boldsymbol{E}$ are $(N \times p)$, $\boldsymbol{X}\boldsymbol{B}$ is $(N \times q \times p)$ and $\text{row}_i(\boldsymbol{E})' \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. Testable

hypotheses, $\mathrm{H}_0 : \boldsymbol{\Theta} = \boldsymbol{\Theta}_0$ have $(a \times b)$ $\boldsymbol{\Theta} = \boldsymbol{CBU}$. The unscaled noncentrality is defined as $\boldsymbol{\Delta} = (\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)' \boldsymbol{M}^{-1} (\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)$, $(b \times b)$, such that $\boldsymbol{M} = \boldsymbol{C}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{C}'$, $(a \times a)$, and rank of $\boldsymbol{M}$ is equal to a. The covariance matrix among the transformed (hypothesis) variables is $\boldsymbol{\Sigma}_* = \boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})\boldsymbol{\Upsilon}'$, $(b \times b)$, with $\boldsymbol{\Upsilon}$, $(b \times b)$, the eigenvectors of $\boldsymbol{\Sigma}_*$, such that $\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}' = \boldsymbol{\Upsilon}'\boldsymbol{\Upsilon} = \boldsymbol{I}_b$ and $\boldsymbol{\lambda}$ the vector of eigenvalues, λ_i, for $\boldsymbol{\Sigma}_*$. If $\nu_e = N - \mathrm{rank}(\boldsymbol{X})$ are the error degrees of freedom between subjects, estimators are $\widehat{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ (when applicable, else $\widetilde{\boldsymbol{B}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$) and $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{Y}'[\boldsymbol{I} - (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}']\boldsymbol{Y}/\nu_e$. Here, $\widehat{\boldsymbol{\Sigma}}$ is the unbiased restricted maximum likelihood (REML) estimator while $\widetilde{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}\nu_e/N$ is the (biased) maximum likelihood estimator (MLE).

The Box conservative, the Geisser-Greenhouse, the Huynh-Feldt, and the uncorrected UNIREP tests may be computed in terms of the estimated hypothesis sums of squares,

$$\widehat{\boldsymbol{\Delta}} = (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)' \boldsymbol{M}^{-1} (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \sim \mathcal{W}_b(a, \boldsymbol{\Sigma}_*, \boldsymbol{\Omega}) \,, \tag{37}$$

and the estimated covariance among the transformed (hypothesis) variables, $\widehat{\boldsymbol{\Sigma}}_* = \boldsymbol{U}'\widehat{\boldsymbol{\Sigma}}\boldsymbol{U}$, with $\boldsymbol{S} = \nu_e \widehat{\boldsymbol{\Sigma}}_* \sim \mathcal{W}_b(\nu_e, \boldsymbol{\Sigma}_*)$. All use the same test statistic,

$$T_u = \frac{\mathrm{tr}(\widehat{\boldsymbol{\Delta}})/a}{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)} \,. \tag{38}$$

Power analysis involves $\boldsymbol{\omega}_* = \{\omega_{*kk}\}$,

$$\omega_{*kk} = \boldsymbol{v}_k' \boldsymbol{\Delta} \boldsymbol{v}_k / \lambda_k \,, \tag{39}$$

the diagonal elements of the scaled noncentrality, $\boldsymbol{\Omega}_* = \boldsymbol{\Upsilon}'\boldsymbol{\Delta}\boldsymbol{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})^{-1} = \boldsymbol{\Delta}_*\mathrm{Dg}(\boldsymbol{\lambda})^{-1}$, such that $\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}' = \boldsymbol{\Upsilon}'\boldsymbol{\Upsilon} = \boldsymbol{I}_b$. Necessarily, $\boldsymbol{\Delta} = \boldsymbol{\Delta}'$ is non-negative definite. Hence $\boldsymbol{\Delta} = \boldsymbol{\Phi}_{\boldsymbol{\Delta}}\boldsymbol{\Phi}_{\boldsymbol{\Delta}}'$, with $\boldsymbol{\Phi}_{\boldsymbol{\Delta}}$ $(b \times s_*)$ for $s_* = \mathrm{rank}(\boldsymbol{\Delta}) = \mathrm{rank}(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)$.

Muller and Barton (1989) derived the exact distribution of $\mathrm{tr}(\widehat{\boldsymbol{\Delta}})$ under the alternative. With the independence of $\widehat{\boldsymbol{\Delta}}$ and $\widehat{\boldsymbol{\Sigma}}_*$, their result allows expressing the test statistic in terms of the independent set $\{y_{kh}, y_{ke}\}$. Here, $y_{kh} \sim \chi^2(a, \omega_{*kk})$ and $y_{ke} \sim \chi^2(\nu_e)$, such that

$$\text{tr}(\widehat{\boldsymbol{\Delta}}) = \sum_{k=1}^{b} \lambda_k y_{kh} = Q(\boldsymbol{\lambda}, a\mathbf{1}_b, \boldsymbol{\omega}_*) \tag{40}$$

and

$$\nu_e \text{tr}(\widehat{\boldsymbol{\Sigma}}_*) = \sum_{k=1}^{b} \lambda_k y_{ke} = Q(\boldsymbol{\lambda}, ab\mathbf{1}_b, \mathbf{0}) \, . \tag{41}$$

Muller *et al.* (2007) showed that the CDF of the UNIREP test statistic could be expressed exactly in terms of a CDF of the sum of b positively and b negatively weighted independent chi-squares,

$$\begin{aligned} \Pr\{T_u \leq t\} = \Pr\left\{ \frac{\text{tr}(\widehat{\boldsymbol{\Delta}})/a}{\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)} \leq t \right\} &= \Pr\left\{ \sum_{k=1}^{b} \lambda_k y_{kh} - (ta/\nu_e) \sum_{k=1}^{b} \lambda_k y_{ke} \leq 0 \right\} \\ &= \Pr\{Q_d[\boldsymbol{\lambda}_d(t), \boldsymbol{\nu}_d, \boldsymbol{\omega}_d] \leq 0\} \, . \end{aligned} \tag{42}$$

Here, $\boldsymbol{\lambda}_d(t) = [\, \boldsymbol{\lambda}' \; -(ta/\nu_e)\boldsymbol{\lambda}' \,]'$, $\boldsymbol{\nu}_d = [\, a\mathbf{1}_b' \; \nu_e \mathbf{1}_b' \,]'$, and $\boldsymbol{\omega}_d = [\, \boldsymbol{\omega}_*' \; \mathbf{0}_b' \,]'$. Their work allows for the computation of exact test size and power for the uncorrected and the Box conservative tests with Davies' (1980) algorithm.

Muller *et al.* (2007) further showed that F approximations work well for the Geisser-Greenhouse and the Huynh-Feldt tests under the null. Using a theorem from Kim *et al.* (2006), they separately matched two noncentral moments and the first central moment of the numerator of the test statistic to a scaled noncentral chi-square, and two moments of the denominator to a scaled central chi-square. This theorem allowed Muller *et al.* (2007) to approximate the UNIREP test statistic with a noncentral F distribution,

$$\Pr\{T_u \leq t\} \approx \Pr\left\{ \frac{\lambda_{*1} y_{*1}/(ab)}{\lambda_{*2} y_{*2}/(b\nu_e)} \leq t \right\} = F_F\left(t \frac{\lambda_{*2}}{\lambda_{*1}} \frac{ab}{\nu_{*1}} \frac{\nu_{*2}}{b\nu_e}; \nu_{*1}, \nu_{*2}, \omega_* \right) . \tag{43}$$

Here, $y_{*1} \sim \chi^2(\nu_{*1}, \omega_*)$, $y_{*2} \sim \chi^2(\nu_{*2})$, $\text{tr}(\widehat{\boldsymbol{\Delta}}) \approx \lambda_{*1} y_{*1}$ and $\text{tr}(\widehat{\boldsymbol{\Sigma}}_*) \approx \lambda_{*2} y_{*2}$.

When sphericity is not met, Box (1954a, b) showed that the UNIREP test statistic under the null could be approximately distributed as an F with reduced degrees of freedom, $F(ab\epsilon, b\nu_e\epsilon)$. The reduction comes from various estimators of ϵ, a measure of sphericity,

used as degrees of freedom multipliers. The four tests differ only by their choice for ϵ, and thus by their respective degrees of freedom. Geisser and Greenhouse (1958) observed that ϵ is bounded, such that $1/b \leq \epsilon \leq 1$. If sphericity exists in the model, $\epsilon = 1$, then $\text{Dg}(\boldsymbol{\lambda}) = \lambda_* \boldsymbol{I}_b$ and $T_u \sim F[ab, b\nu_e, \text{tr}(\boldsymbol{\Omega})]$ exactly. Under sphericity, the test is exactly size alpha and uniformly most powerful among similarly invariant tests. Sphericity estimates are always ordered Box conservative (Box), Geisser-Greenhouse (GG), Huynh-Feldt (HF), and uncorrected (UN); specifically $1/b \leq \widehat{\epsilon} \leq \widetilde{\epsilon} \leq 1$. Test size and power are in the same order.

The Geisser-Greenhouse test is based on the MLE,

$$\widehat{\epsilon} = \frac{\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)}{b\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)} = \frac{\text{tr}^2(\widetilde{\boldsymbol{\Sigma}}_*)}{b\text{tr}(\widetilde{\boldsymbol{\Sigma}}_*^2)} = \frac{\text{tr}^2(\boldsymbol{S})}{b\text{tr}(\boldsymbol{S}^2)} \,, \tag{44}$$

while the Huynh-Feldt test uses an approximately unbiased estimator (a 1-1 function of the MLE),

$$\widetilde{\epsilon} = \frac{N b\widehat{\epsilon} - 2}{b(\nu_e - b\widehat{\epsilon})} \,. \tag{45}$$

In practice, the Huynh-Feldt estimator is truncated above at 1.0.

Test T with multiplier ϵ_T has critical value $t_{\text{crit}}(\text{T}) = F_F^{-1}(1-\alpha; \epsilon_T \cdot ab, \epsilon_T \cdot b\nu_e)$. Thus, the associated power approximations use

$$t(\text{Un}) = F_F^{-1}[1-\alpha; \quad 1 \cdot ab, \quad 1 \cdot b\nu_e] \tag{46}$$
$$t(\text{HF}) = F_F^{-1}[1-\alpha; \text{E}(\widetilde{\epsilon}) \cdot ab, \text{E}(\widetilde{\epsilon}) \cdot b\nu_e] \tag{47}$$
$$t(\text{GG}) = F_F^{-1}[1-\alpha; \text{E}(\widehat{\epsilon}) \cdot ab, \text{E}(\widehat{\epsilon}) \cdot b\nu_e] \tag{48}$$
$$t(\text{Box}) = F_F^{-1}[1-\alpha; \ 1/b \cdot ab, \ 1/b \cdot b\nu_e] \,. \tag{49}$$

Here, $\text{E}(\widetilde{\epsilon})$ and $\text{E}(\widehat{\epsilon})$ are approximate, expected sphericity estimator values derived by Muller *et al.* (2007).

The Box conservative and uncorrected tests have constant critical values, while random multipliers, $\widehat{\epsilon}$ and $\widetilde{\epsilon}$, yield random critical values for the Geisser-Greenhouse and

Huynh-Feldt tests. Muller and Barton (1989) proposed accurate power approximations for all four UNIREP tests using a noncentral F distribution. Muller *et al.* (2007) expanded upon the work of Muller and Barton (1989), and presented their own power approximations, which consistently performed as well as, or better than, those of Muller and Barton (1989). Only the Muller *et al.* (2007) approximations will be considered for the remainder of this discussion.

For known covariance and means, the Muller *et al.* (2007) UNIREP power approximations are all of the form

$$P = 1 - F_F\left[F_F^{-1}(1-\alpha; e_1 \cdot ab, e_2 \cdot b\nu_e); e_3 \cdot ab, e_4 \cdot \nu_e b, \frac{\text{tr}(\boldsymbol{\Delta})}{\overline{\lambda}/e_5}\right]. \tag{50}$$

Here, $\overline{\lambda}$ is equal to $\text{tr}(\boldsymbol{\Sigma}_*)/b$ with b equal to the rank of $\boldsymbol{\Sigma}_*$. The parameters e_1 through e_5 represent various sphericity values used at certain points in the power approximation. They are derived in Appendix A of Muller *et al.* (2007). The parameter ϵ_n is the sphericity parameter under the nonnull case,

$$\epsilon_n = \frac{\text{tr}^2(\boldsymbol{\Sigma}_*) + 2\text{tr}(\boldsymbol{\Sigma}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b[\text{tr}(\boldsymbol{\Sigma}_*^2) + 2\text{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta}/a)]} \tag{51}$$

$$\epsilon_d = \frac{\text{tr}^2(\boldsymbol{\Sigma}_*)}{b\text{tr}(\boldsymbol{\Sigma}_*^2)} = \epsilon_n|\boldsymbol{\Delta} = \mathbf{0} \qquad \equiv \epsilon\,. \tag{52}$$

As depicted above in equation 52, the parameter ϵ_n reduces to the familiar sphericity parameter under the null case. In Table 4.1, particular values are summarized for e_1 through e_5 for the four UNIREP tests, when both covariance and means are known.

Table 4.1. Sphericity Multipliers for UNIREP Power Approximations for $\boldsymbol{\Sigma}_*$, $\boldsymbol{\Delta}$ (Both Known)

Test	e_1	e_2	e_3	e_4	e_5
Un	1	1	ϵ_n	ϵ_d	ϵ_n
HF	$\text{E}(\widetilde{\epsilon})$	$\text{E}(\widetilde{\epsilon})$	ϵ_n	ϵ_d	ϵ_n
GG	$\text{E}(\widehat{\epsilon})$	$\text{E}(\widehat{\epsilon})$	ϵ_n	ϵ_d	ϵ_n
Box	$1/b$	$1/b$	ϵ_n	ϵ_d	ϵ_n

4.3 Population Properties of UNIREP Power Approximations in terms of known $\boldsymbol{\Sigma}_*$ and $\boldsymbol{\Delta}$

The Huynh-Feldt estimator, $\widetilde{\epsilon} = (N\widehat{b\epsilon} - 2)/[b(\nu_e - \widehat{b\epsilon})]$, described by Huynh and Feldt (1976), was incorrectly proposed as the ratio of two unbiased estimators. Their claim holds true only for the special case of rank of $\boldsymbol{X}$ equal to 1. In section 2.3, a rank-adjusted approximately unbiased estimator was introduced and shown to better estimate the population sphericity parameter than the Huynh-Feldt estimator when rank of $\boldsymbol{X}$ was greater than 1. This result was observed in all cases except sphericity ($\epsilon = 1$). In the case of sphericity, neither the Huynh-Feldt nor the rank-adjusted tests would be considered. When rank of $\boldsymbol{X}$ is equal to 1, the rank-adjusted sphericity estimator reduces to that of the Huynh-Feldt sphericity estimator. The rank-adjusted estimator achieves what Huynh and Feldt (1976) had originally intended to accomplish. The rank-adjusted estimator is a ratio of two unbiased estimators, which is less biased and less dependent on large sample sizes than the Geisser-Greenhouse estimator when the covariance matrix deviates only moderately from sphericity, and is so for any rank of $\boldsymbol{X}$. The rank-adjusted estimator is

$$\widetilde{\epsilon}_r = \frac{(\nu_e + 1)\widehat{b\epsilon} - 2}{b(\nu_e - \widehat{b\epsilon})} \,. \tag{53}$$

Like the Huynh-Feldt estimator, the rank-adjusted estimator is truncated to 1.0 when necessary. The rank-adjusted estimator will always be less than or equal to the Huynh-Feldt estimator, with equality only if rank of $\boldsymbol{X}$ is equal to 1.

In section 3.3, a rank-adjusted power approximation was proposed. The approximation was similar to that of the Huynh-Feldt power approximation presented in Muller *et al.* (2007). The rank-adjusted power approximation was shown through simulations to approximate observed mean power values as well as, or better than, the Huynh-Feldt power approximation for practical research purposes. Furthermore, the rank-adjusted power approximation was shown to adequately control test size when rank of $\boldsymbol{X}$

was greater than 1. Meanwhile, in some cases, the Huynh-Feldt test size was double the target test size. For the remainder of this paper, the rank-adjusted Huynh-Feldt estimator, test and power approximation will be used in place of the Huynh-Feldt estimator, test and power approximation, respectively.

Muller *et al.* (2007) showed that if $S_{t1} = \sum_{k=1}^{b} \lambda_k$, $S_{t2} = \sum_{k=1}^{b} \lambda_k^2$, $S_{t3} = \sum_{k=1}^{b} \lambda_k \omega_{*kk}$ and $S_{t4} = \sum_{k=1}^{b} \lambda_k^2 \omega_{*kk}$, then

$$\lambda_{*1} = \frac{(aS_{t2} + 2S_{t4})}{(aS_{t1} + 2S_{t3})} \tag{54}$$

$$\nu_{*1} = aS_{t1}/\lambda_{*1} \tag{55}$$

$$\omega_* = S_{t3}/\lambda_{*1} \tag{56}$$

$$\lambda_{*2} = S_{t2}/S_{t1} \tag{57}$$

$$\nu_{*2} = \nu_e S_{t1}^2/S_{t2} = \nu_e b\epsilon \; . \tag{58}$$

They used these parameters to approximate the UNIREP test statistic with a noncentral F distribution, as presented in equation 43.

Lemma 4.1 Defining

$$S_{t1} = \sum_{k=1}^{b} \lambda_k \; , \quad S_{t2} = \sum_{k=1}^{b} \lambda_k^2 \; , \quad S_{t3} = \sum_{k=1}^{b} \lambda_k \omega_{*kk} \; , \quad S_{t4} = \sum_{k=1}^{b} \lambda_k^2 \omega_{*kk} \tag{59}$$

implies $\{S_{t1}, S_{t2}, S_{t3}, S_{t4}\} = \{\text{tr}(\boldsymbol{\Sigma}_*), \text{tr}(\boldsymbol{\Sigma}_*^2), \text{tr}(\boldsymbol{\Delta}), \text{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta})\}$.

Lemma 4.2 The constant in the critical value of the UNIREP test statistic approximation introduced by Muller *et al.* (2007) is equal to 1,

$$\frac{\lambda_{*2}}{\lambda_{*1}} \frac{ab}{\nu_{*1}} \frac{\nu_{*2}}{b\nu_e} = 1 \; . \tag{60}$$

Thus,

$$\Pr\{T_u \leq t\} \approx \Pr\left\{ \frac{\lambda_{*1} y_{*1}/(ab)}{\lambda_{*2} y_{*2}/(b\nu_e)} \leq t \right\} = F_F(t; \nu_{*1}, \nu_{*2}, \omega_*) \; . \tag{61}$$

4.4 Estimated Properties of UNIREP Power Approximations as a function of $\widehat{\boldsymbol{\Sigma}}_*$ and $\boldsymbol{\Delta}$

For some purposes, various collections of elements of $\{e_1, e_2, e_3, e_4, e_5, \text{tr}(\boldsymbol{\Delta}), \overline{\lambda}\}$ are estimates of (approximate) means or quantiles, and hence random. These random elements imply random power. In the case of estimated covariance and fixed means, $\{\widehat{\boldsymbol{\Sigma}}_*, \boldsymbol{\Delta}\}$, $\widehat{\boldsymbol{\Sigma}}_* = \widehat{\boldsymbol{E}}'\widehat{\boldsymbol{E}}/\nu_{\text{est}}$ indicates the REML (unbiased) estimator. A distinction must be carefully maintained between the estimation study and target study. The estimation study provides the covariance estimate and has sample size, rank of $\boldsymbol{X}$ and degrees of freedom, N_{est}, rank$(\boldsymbol{X}_{\text{est}})$ and $\nu_{\text{est}} = N_{\text{est}} - \text{rank}(\boldsymbol{X}_{\text{est}})$, respectively. The target study for which power is desired has sample size, rank of $\boldsymbol{X}$ and degrees of freedom, N, rank$(\boldsymbol{X})$ and $\nu_e = N - \text{rank}(\boldsymbol{X})$, respectively.

The parameter $\widehat{\epsilon}_n$ is the estimated sphericity parameter under the nonnull case, which reduces to the familiar estimated sphericity parameter, $\widehat{\epsilon}$, under the null case,

$$\widehat{\epsilon}_n = \frac{\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\Big[\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\Big]} \tag{62}$$

$$\widehat{\epsilon} \equiv \widehat{\epsilon}_d = \frac{\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)}{b\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)} = \widehat{\epsilon}_n|\boldsymbol{\Delta} = \boldsymbol{0}\,. \tag{63}$$

Lemma 4.3 For the nonnull case, a ratio estimating ϵ_n in terms of correlated, but unbiased, estimators is

$$\widetilde{\epsilon}_n = \frac{\nu_{\text{est}}(\nu_{\text{est}}+1)\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2\nu_{\text{est}}\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) + 2[\nu_{\text{est}}(\nu_{\text{est}}+1) - 2]\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\Big\{\nu_{\text{est}}^2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_{\text{est}}\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) + 2[\nu_{\text{est}}(\nu_{\text{est}}+1) - 2]\text{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\Big\}}\,. \tag{64}$$

$$\widetilde{\epsilon}_r = \frac{(\nu_{\text{est}}+1)b\widehat{\epsilon} - 2}{b(\nu_{\text{est}} - b\widehat{\epsilon})} = \widetilde{\epsilon}_n|\boldsymbol{\Delta} = \boldsymbol{0}\,. \tag{65}$$

This approximately unbiased estimator reduces to the rank-adjusted Huynh-Feldt sphericity estimator under the null case, as depicted above in equation 65.

For estimated covariance and fixed means, estimated UNIREP power approximations are all of the form

$$P = 1 - F_F\left[F_F^{-1}(1-\alpha; e_1 \cdot ab, e_2 \cdot b\nu_e); e_3 \cdot ab, e_4 \cdot \nu_e b, \frac{\text{tr}(\boldsymbol{\Delta})}{\overline{\widehat{\lambda}}/e_5}\right], \tag{66}$$

with $\overline{\widehat{\lambda}}$ equal to $\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)/b$. In Table 4.2, particular values are summarized for e_1 through e_5 for the four UNIREP tests. The values for e_1 and e_2 are natural choices for the various UNIREP tests. The particular values of e_3, e_4 and e_5 were chosen based on the results of extensive, experimental simulations. Nearly every combination of $\widehat{\epsilon}_n, \widetilde{\epsilon}_n, \widehat{\epsilon}_d, \widetilde{\epsilon}_r$, 1 and $1/b$ was examined thoroughly for each UNIREP test for the wide range of simulations discussed in Muller *et al.* (2007). The values chosen provided the most accurate results. In retrospect, they are natural choices as well.

Table 4.2. Sphericity Multipliers for Approximately Unbiased UNIREP Power Approximations as a function of $\widehat{\boldsymbol{\Sigma}}_*$, $\boldsymbol{\Delta}$ (Estimated Covariance)

Test	e_1	e_2	e_3	e_4	e_5
Un	1	1	$\widetilde{\epsilon}_n$	$\widehat{\epsilon}_d$	$\widetilde{\epsilon}_n$
HF	$\widetilde{\epsilon}_r$	$\widetilde{\epsilon}_r$	$\widetilde{\epsilon}_n$	$\widehat{\epsilon}_d$	$\widetilde{\epsilon}_n$
GG	$\widehat{\epsilon}_d$	$\widehat{\epsilon}_d$	$\widetilde{\epsilon}_n$	$\widehat{\epsilon}_d$	$\widetilde{\epsilon}_n$
Box	$1/b$	$1/b$	$\widetilde{\epsilon}_n$	$\widehat{\epsilon}_d$	$\widetilde{\epsilon}_n$

4.5 Approximate Power Confidence Intervals for UNIREP Tests

Taylor and Muller (1995) demonstrated that a function of the variance parameter exactly follows a chi-square distribution. They developed exact bounds for the noncentrality, ω, in the univariate setting by realizing that ω is an inverse function of the variance parameter, $\omega = \delta/\sigma^2$, such that $\delta = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\boldsymbol{M}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = S_H(\boldsymbol{\theta}, N)$. Thus, exact confidence intervals may be calculated for power as well, due to the strict monotone dependence of the noncentral F distribution function on the noncentrality, specifically, $\Pr(\widehat{\omega}_L \leq \omega \leq \widehat{\omega}_U) = \Pr(\widehat{P}_L \leq P \leq \widehat{P}_U)$.

Here, the UNIREP problem is approached in much the same way that Taylor and Muller (1995) did in the univariate case. The Muller *et al.* (2007) power approximations for known covariance and means always defines e_5 to be ϵ_n. Thus, the noncentrality is ω_*, such that $\omega_* = [\text{tr}(\mathbf{\Delta})]/(\overline{\lambda}/\epsilon_n)$.

With

$$\omega_* = \text{tr}(\mathbf{\Delta})/\lambda_{*1} \tag{67}$$

$$\lambda_{*1} = \frac{\text{tr}(\mathbf{\Sigma}_*^2) + 2\text{tr}(\mathbf{\Delta}\mathbf{\Sigma}_*/a)}{\text{tr}(\mathbf{\Sigma}_*) + 2\text{tr}(\mathbf{\Delta}/a)} = \overline{\lambda}/\epsilon_n \tag{68}$$

$$\epsilon_n = \frac{\text{tr}^2(\mathbf{\Sigma}_*) + 2\text{tr}(\mathbf{\Sigma}_*)\text{tr}(\mathbf{\Delta}/a)}{b[\text{tr}(\mathbf{\Sigma}_*^2) + 2\text{tr}(\mathbf{\Sigma}_*\mathbf{\Delta}/a)]}\,, \tag{69}$$

it follows that

$$\omega_* = \text{tr}(\mathbf{\Delta}) \cdot \frac{\text{tr}(\mathbf{\Sigma}_*) + 2\text{tr}(\mathbf{\Delta}/a)}{\text{tr}(\mathbf{\Sigma}_*^2) + 2\text{tr}(\mathbf{\Delta}\mathbf{\Sigma}_*/a)}\,. \tag{70}$$

To stay consistent with the notation presented in Taylor and Muller (1995), note that $\text{tr}(\mathbf{\Delta})$ is, in actuality, the trace of the (population) hypothesis sums of squares. This realization makes λ_{*1}, in a sense, a form of the variance parameter.

For estimated covariance and fixed means, a ratio involving one biased and two unbiased estimators (Appendix B has derivations of the first moments) may be written as

$$\widetilde{\lambda}_{*1} = \frac{\text{tr}(\widehat{\mathbf{\Sigma}}_*^2) + 2\text{tr}(\mathbf{\Delta}\widehat{\mathbf{\Sigma}}_*/a)}{\text{tr}(\widehat{\mathbf{\Sigma}}_*) + 2\text{tr}(\mathbf{\Delta}/a)}\,. \tag{71}$$

In the univariate setting, Taylor and Muller (1995) were forced to deal with only one random variable. In the UNIREP case, there are three, all of which are correlated.

Still, similar to the univariate setting, $\widetilde{\lambda}_{*1}$ can be approximated with a chi-square using a Satterthwaite approximation,

$$\frac{\widetilde{\lambda}_{*1}\nu_*}{\lambda_{*1}} \sim \chi^2(\nu_*)\,. \tag{72}$$

Here, $\nu_* = (b\nu_{\text{est}}) \cdot \widehat{\epsilon}_d/\widetilde{\epsilon}_n$. Lower and upper tail probabilities, α_L and α_U, respectively,

define the confidence coefficient, $p_{CL} = 1 - \alpha_L - \alpha_U$. Also, $c_{\alpha L} = F^{-1}_{\chi^2}(\alpha_L; \nu_*)$ and $c_{\alpha U} = F^{-1}_{\chi^2}(1 - \alpha_U; \nu_*)$ are the α_L and $(1 - \alpha_U)$ quantiles of a central chi-square distribution with ν_* degrees of freedom, respectively. Approximate confidence limits for the noncentrality, and thus power, may be calculated in the UNIREP setting as:

$$\Pr\left\{ c_{\alpha L} < \frac{\widetilde{\lambda}_{*1}\nu_*}{\lambda_{*1}} < c_{\alpha U} \right\} \approx p_{CL} \tag{73}$$

$$\Updownarrow \Pr\left\{ \frac{c_{\alpha L}}{\widetilde{\lambda}_{*1}\nu_*} < \frac{1}{\lambda_{*1}} < \frac{c_{\alpha U}}{\widetilde{\lambda}_{*1}\nu_*} \right\} \approx p_{CL} \tag{74}$$

$$\Updownarrow \Pr\left\{ \frac{\mathrm{tr}(\Delta) c_{\alpha L}}{\widetilde{\lambda}_{*1}\nu_*} < \frac{\mathrm{tr}(\Delta)}{\lambda_{*1}} < \frac{\mathrm{tr}(\Delta) c_{\alpha U}}{\widetilde{\lambda}_{*1}\nu_*} \right\} \approx p_{CL} \tag{75}$$

$$\Updownarrow \Pr\left\{ \frac{\mathrm{tr}(\Delta) c_{\alpha L}}{\widetilde{\lambda}_{*1}\nu_*} < \omega_* < \frac{\mathrm{tr}(\Delta) c_{\alpha U}}{\widetilde{\lambda}_{*1}\nu_*} \right\} \approx p_{CL} \tag{76}$$

$$\Updownarrow \Pr\left\{ g\left(\frac{\mathrm{tr}(\Delta) c_{\alpha L}}{\widetilde{\lambda}_{*1}\nu_*} \right) < \text{Power} < g\left(\frac{\mathrm{tr}(\Delta) c_{\alpha U}}{\widetilde{\lambda}_{*1}\nu_*} \right) \right\} \approx p_{CL} \, . \tag{77}$$

Approximate lower and upper bounds on the noncentrality, $\widetilde{\omega}_{*L}$ and $\widetilde{\omega}_{*U}$, respectively, are thus defined as

$$\widetilde{\omega}_{*L} = \frac{\mathrm{tr}(\Delta) c_{\alpha L}}{\widetilde{\lambda}_{*1}\nu_*} \tag{78}$$

and

$$\widetilde{\omega}_{*U} = \frac{\mathrm{tr}(\Delta) c_{\alpha U}}{\widetilde{\lambda}_{*1}\nu_*} \, . \tag{79}$$

The strict monotone dependence of the noncentral F function on the approximate noncentrality ensures an approximate confidence interval for power. Here, lower and upper bounds on power, $\widetilde{P}_L$ and $\widetilde{P}_U$, respectively, are defined as

$$\widetilde{P}_L = 1 - F_F\left[F_F^{-1}(1 - \alpha; e_1 \cdot ab, e_2 \cdot b\nu_e); e_3 \cdot ab, e_4 \cdot \nu_e b, \widetilde{\omega}_{*L} \right] \tag{80}$$

and

$$\widetilde{P}_U = 1 - F_F\left[F_F^{-1}(1-\alpha; e_1 \cdot ab, e_2 \cdot b\nu_e); e_3 \cdot ab, e_4 \cdot \nu_e b, \widetilde{\omega}_{*U}\right], \qquad (81)$$

such that e_1 through e_4 are defined in Table 4.2 above.

4.6 Simulations

The accuracy of the new approximate confidence intervals for UNIREP tests in the case of an estimated covariance and fixed means were evaluated for a wide range set of simulations. In section 4.6.1, the new technique was applied to the CLAHE mammography example presented in Muller *et al.* (2007). The CLAHE mammography example illustrates a completely within-subject design. In section 4.6.2, the technique is applied to an example that tests the interaction of orthonormal trends for between- and within-subjects. The latter example provides a larger number of conditions that may affect the accuracy of the coverage. In particular, the effect of modifying the estimating study is examined.

Appendix C contains a more detailed description of the simulation conditions and the examples. All simulations were conducted in SAS/IML (SAS 9.1, SAS Institute, Copyright 2003). Simulated observed mean power values were computed using a modified version of LINMOD 3.3 (http://ehpr.ufl.edu/muller/), a software that performs a wide variety of General Linear Multivariate Model computations. LINMOD 3.3 was modified to include the rank-adjusted Huynh-Feldt estimator and test. Predicted power values and approximate power confidence intervals were computed using a modified version of POWERLIB 2.0 (http://ehpr.ufl.edu/muller/), a software that computes statistical power for the General Linear Multivariate Model. POWERLIB 2.0 was modified to include the rank-adjusted test. The modified versions of both LINMOD and POWERLIB will be made available soon.

4.6.1 Simulation 1

The accuracy of the new approximate confidence intervals for UNIREP tests in the case of an estimated covariance and fixed means were evaluated for a completely within-subject design with 9 repeated measures and sample sizes of 10, 20 and 40,

$$\underset{(N \times 9)}{\boldsymbol{Y}} = \underset{(N \times 1 \times 9)}{\mathbf{1}_N \boldsymbol{B}} + \underset{(N \times 9)}{\boldsymbol{E}} \quad . \tag{82}$$

Appropriate fixed matrices of regression coefficients, $\boldsymbol{B}$, contrast matrices, $\boldsymbol{C}$ and $\boldsymbol{U}$, and $\boldsymbol{\Theta}_0$ were chosen to test a within-subject interaction for a test size, α, of 0.05. The matrices were also chosen to ensure approximate target predicted power values for the Geisser-Greenhouse test of 0.20, 0.50 and 0.80, using the power approximation introduced in Muller *et al.* (2007). Specific design matrices, $\boldsymbol{X}$, were defined. Population covariance matrices were chosen to provide specific population sphericity values, $\epsilon \in \{0.282, 0.505, 0.720, 1.00\}$. Pseudo-random realizations of the error matrix, $\boldsymbol{E}$, were generated and appropriate test statistics were calculated. The observed mean power values for the four UNIREP tests were calculated and tabulated for $500{,}000$ replications per condition. The observed mean power value for each condition was the proportion of rejected tests among $500{,}000$ simulated realizations.

For the conditions described above, additional pseudo-random realizations of the error matrix were generated using an estimating study with sample size, N_{est}, of 10 and rank of $\boldsymbol{X}$, rank$(\boldsymbol{X}_{\text{est}})$, of 1 with $500{,}000$ replications per condition for all four UNIREP tests. Corresponding estimated covariance matrices were calculated, as well as lower and upper bounds for power using the methods presented in section 4.5. Approximate confidence interval coverage was defined as the proportion of the $500{,}000$ simulated bound realizations that successfully covered the observed mean power values for each condition described above. This number of replications was chosen to ensure a standard error of observed mean estimates less than or equal to 0.0003, nearly guaranteeing 3 digits of accuracy. Only coverage of observed mean power values, and not predicted, was tabulated. The accuracy of the predicted power values, with respect to the observed, made it essentially redundant to consider both. Both one- and two-sided confidence intervals were evaluated with target coverages of 90% and 95%.

Only the worst case results for two-sided 95% confidence intervals are presented here. The worst cases occurred with the smallest sample size for the target study. In Table 4.3, the proportion of simulations in which the estimated confidence interval successfully covered the observed mean population power is shown for the Box conservative test. The results presented are for a target sample size of 10, for the four population sphericity values and three target power values considered. The lower tail value is specifically the lower error, tabulating the proportion of simulations in which the approximate confidence interval fell below the observed mean population power. A similar proportion was tabulated for the upper tail value. For the Box conservative test, for a wide range of population sphericity values and target power values, the target 95% estimated coverage is consistently reached. There are two cases in which the target coverage is not reached, and they both occur in cases with large population sphericity values. In both cases, the observed mean population powers are extremely low. Under these conditions, the Box conservative test would not be used in practice.

Tables 4.4-4.5 contain similar results for the Geisser-Greenhouse and the Huynh-Feldt tests, respectively. The target 95% estimated coverage is consistently reached in the cases of extreme population sphericity values for the Geisser-Greenhouse and Huynh-Feldt tests. For midrange population sphericity values, the approximated coverage fell below the target coverage by as little as 0.8% and as much as 7.3% for the Geisser-Greenhouse, and as little as 1.4% and as much as 12.1% for the Huynh-Feldt. Coverage accuracy seemed to improve as the estimated population power increased. In practical biomedical research, lower power values are of little concern. Rarely will one have a power analyses targeted below 0.70. For the highest target power value, 0.80, the largest deviation from the target 95% estimated coverage was 2.6% and 4.1% for the Geisser-Greenhouse and Huynh-Feldt tests, respectively. Both occurred for the population sphericity value of 0.505.

In the case of the uncorrected test, only a spherical case need be considered. When sphericity is met, the uncorrected test is the uniformly most powerful exact size alpha test, among similarly invariant tests. If sphericity is not met, a corrected test such as the Geisser-Greenhouse, the Huynh-Feldt or the Box conservative would be more appropriate to use. In Table 4.6, the proportion of simulated realizations in which the estimated confidence interval successfully covered the observed mean population power is shown for the uncorrected test for all sample sizes considered with a spherical population. The approximation always reached the target estimated coverage for the uncorrected test.

Realizing that exact confidence intervals exist in the case of sphericity is important. Achieving the exact results require using the correct (maximum likelihood) estimates for the common variance and covariance (Morrison, 1990), rather than the unstructured covariance estimate used in the power program, POWERLIB. Additional details are in the POWERLIB manual, and are mostly associated with degrees of freedom corresponding to making all choices of e_1 through e_5 equal to 1.

Although not presented here, in general, the conservative coverage values observed for the Box conservative and the uncorrected tests slowly approached the target coverage value as target sample size increased. This trend was also observed for the conservative coverage values for the extreme population sphericity values for the Geisser-Greenhouse and the Huynh-Feldt tests. The same is true of the liberal coverage values observed for the midrange population sphericity values for the Geisser-Greenhouse and the Huynh-Feldt tests. Similar results were obtained for the target 90% two-sided confidence interval coverage, as well as the 95% and 90% one-sided confidence intervals coverage.

Table 4.3. Target 95% CI (Two-Sided) Estimated Coverage (× 100) of Simulated Population Powers (× 100) for the Box Conservative ($N = 10$), 95% Half Confidence Interval is 6.04×10^{-4}

ϵ	Population Power	Lower Tail	Coverage	Upper Tail
0.282	12.3	1.1	97.8	1.1
	53.5	1.9	97.0	1.1
	93.0	1.7	97.3	1.0
0.505	05.4	0.1	97.3	2.6
	26.6	0.5	97.0	2.5
	69.0	1.1	97.0	1.9
0.720	05.2	0.4	94.1	5.5
	22.7	0.6	96.8	2.6
	56.9	1.4	97.0	1.6
1	02.3	0.6	85.1	14.3
	11.7	0.5	96.0	3.5
	35.0	0.8	97.8	1.4

Table 4.4. Target 95% CI (Two-Sided) Estimated Coverage (× 100) of Simulated Population Powers (× 100) for the Geisser-Greenhouse ($N = 10$), 95% Half Confidence Interval is 6.04×10^{-4}

ϵ	Population Power	Lower Tail	Coverage	Upper Tail
0.282	15.5	3.1	94.7	2.2
	58.5	2.6	95.6	1.8
	94.2	1.8	96.6	1.6
0.505	16.2	5.4	87.7	6.9
	52.0	3.8	90.6	5.6
	87.0	2.6	92.4	5.0
0.720	20.3	2.4	92.3	5.3
	53.9	2.6	94.1	3.3
	85.6	3.3	94.2	2.5
1	16.1	0.7	95.6	3.7
	43.8	1.4	97.0	1.6
	75.1	2.7	96.2	1.1

Table 4.5. Target 95% CI (Two-Sided) Estimated Coverage (× 100) of Simulated Population Powers (× 100) for the Huynh-Feldt ($N = 10$), 95% Half Confidence Interval is 6.04×10^{-4}

ϵ	Population Power	Lower Tail	Coverage	Upper Tail
0.282	16.6	3.8	93.5	2.7
	60.2	2.8	95.2	2.0
	94.6	1.9	96.3	1.8
0.505	21.0	8.2	82.9	8.9
	59.2	4.7	88.5	6.8
	90.2	2.9	90.9	6.2
0.720	27.1	3.6	90.9	5.5
	63.1	3.4	93.3	3.3
	90.4	4.0	93.6	2.4
1	22.4	0.8	96.7	2.5
	53.1	1.8	97.1	1.1
	82.1	3.2	95.9	0.9

Table 4.6. Target 95% CI (Two-Sided) Estimated Coverage (× 100) of Simulated Population Powers (× 100) for the Uncorrected ($\epsilon = 1.00$), 95% Half Confidence Interval is 6.04×10^{-4}

N	Population Power	Lower Tail	Coverage	Upper Tail
10	23.8	0.5	97.5	2.0
	55.1	1.5	97.6	0.9
	83.5	3.2	96.1	0.7
20	21.5	0.8	97.3	1.9
	52.0	1.6	97.6	0.8
	81.4	3.1	96.2	0.7
40	20.7	0.9	97.2	1.9
	50.9	1.5	97.7	0.8
	80.6	3.0	96.3	0.7

The methods presented in section 4.5 allow calculating more than confidence intervals for a single power value at a time. The logic of a proof described in Taylor and Muller (1995) guarantees that accurate confidence *regions* are provided by the point-wise calculations. Figures 4.1-4.4 give graphical representations of approximate power

confidence regions surrounding predicted power curves for the four UNIREP tests. The approximate confidence regions are shown for the population sphericity values most closely related to the common and most practical use of the respective UNIREP tests.

The value of a graphical representation such as Figure 4.1 for the Box conservative test is for researchers to realize that they should be extremely cautious about power analysis results. For the example examined in Figure 4.1, it seems that one may do just as well guessing at the power analysis results. In turn, a benefit of such results may be to provide evidence for the need of an internal pilot study design, which calls for the reevaluation of the power analysis once a portion of data has been collected. Another benefit may be for researchers to suggest the creation of only one-sided power confidence intervals.

Taylor and Muller (1995) suggested using one-sided power confidence intervals in the univariate case claiming that "the change from a one-sided to a two-sided confidence interval has little effect on the upper bound, but a large effect on the lower bound." Muller and Fetterman (2002) provided examples showing that use of a one-sided power confidence interval in the univariate case resulted in tighter and more informative bounds when compared to the two-sided confidence intervals.

The confidence bounds for all four UNIREP tests seemed to converge quickly to the predicted power curve as the population sphericity increased, as evidenced by additional figures created, but not shown here. This suggests that greater population sphericity is associated with less uncertainty for power.

Figure 4.1. Approximate 95% Confidence Region for Predicted Power of the Box Conservative Test of Interaction over $\mathrm{tr}(\Delta)$ with $N = 10$ and Population $\epsilon = 0.282$ for conditions described in Section *4.6.1 Simulation 1*.

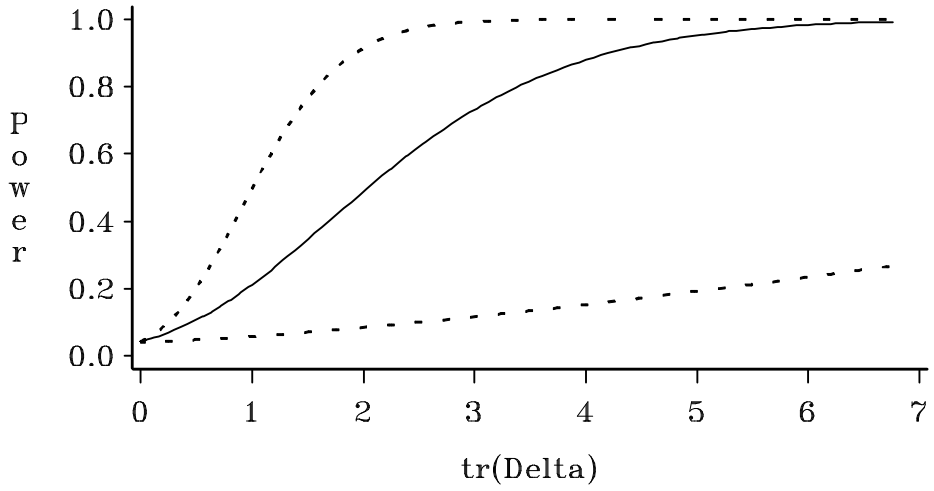


Figure 4.2. Approximate 95% Confidence Region for Predicted Power of the Geisser-Greenhouse Test of Interaction over $\mathrm{tr}(\Delta)$ with $N = 10$ and Population $\epsilon = 0.505$ for conditions described in Section *4.6.1 Simulation 1*.

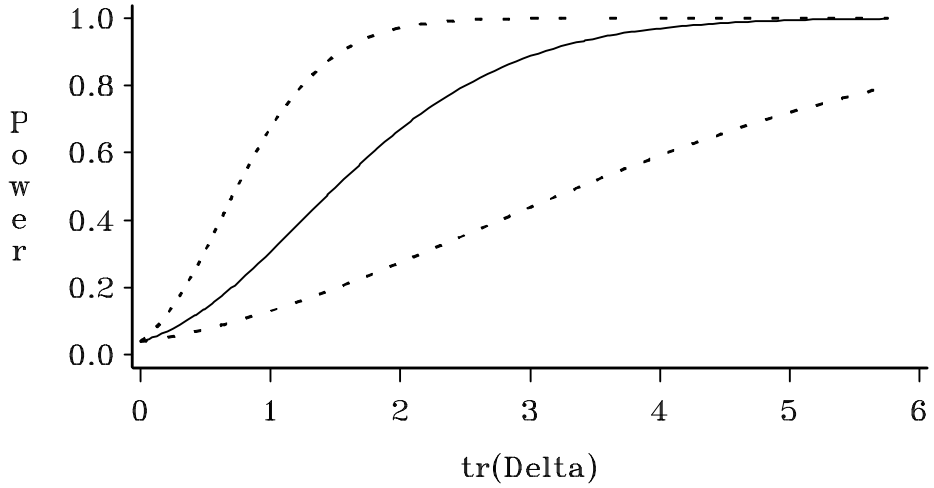


Figure 4.3. Approximate 95% Confidence Region for Predicted Power of the Huynh-Feldt Test of Interaction over tr(Δ) with $N = 10$ and Population $\epsilon = 0.720$ for conditions described in Section *4.6.1 Simulation 1*.

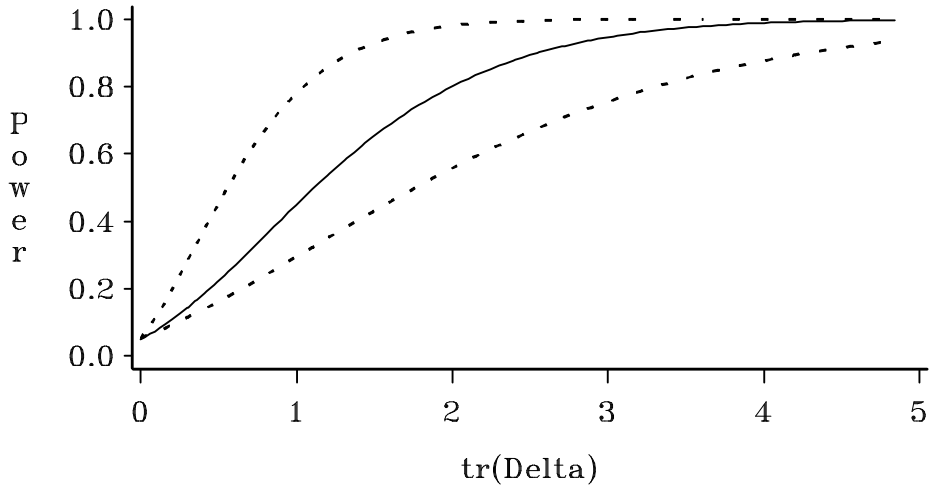
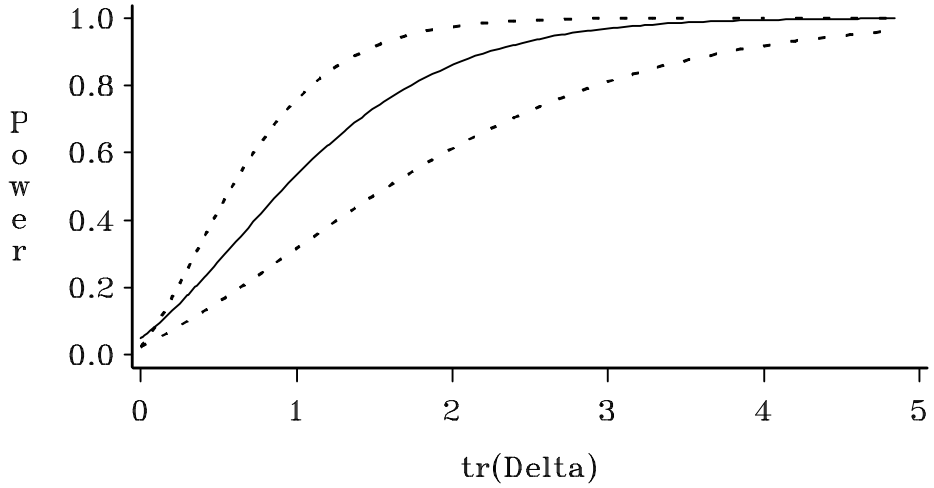


Figure 4.4. Approximate 95% Confidence Region for Predicted Power of the Uncorrected Test of Interaction over tr(Δ) with $N = 10$ and Population $\epsilon = 1.00$ for conditions described in Section *4.6.1 Simulation 1*.



4.6.2 Simulation 2

All of the simulations in the second example considered the condition of rank of $\boldsymbol{X}$ greater than 1. The cases consisted of $p = 5$ repeated measures, N the sample size of 16, 32 and 48, and q the rank of $\boldsymbol{X}$ equal to 4, 8 and 16, in the model

$$\underset{(N \times 5)}{\boldsymbol{Y}} = \underset{(N \times q \times 5)}{\boldsymbol{XB}} + \underset{(N \times 5)}{\boldsymbol{E}} \quad . \tag{83}$$

Appropriate fixed matrices of regression coefficients, $\boldsymbol{B}$, contrast matrices, $\boldsymbol{C}$ and $\boldsymbol{U}$, and $\boldsymbol{\Theta}_0$ were chosen to test a within-subject interaction for a test size, α, of 0.05. The matrices were also chosen to ensure approximate target predicted power values for the rank-adjusted Huynh-Feldt test of 0.20, 0.50 and 0.80, using the power approximation presented in section 3.3. Specific design matrices, $\boldsymbol{X}$, were defined. Population covariance matrices were chosen to provide specific population sphericity values, $\epsilon \in \{0.282, 0.505, 0.720, 1.00\}$. Observed mean power values were simulated and tabulated in a similar manner to that described in section 4.6.1.

Pseudo-random realizations of the error matrix were generated using an estimating study with sample size, N_{est}, of 16 and rank of $\boldsymbol{X}$, rank($\boldsymbol{X}_{\text{est}}$), of 4 with $500{,}000$ replications per condition for all four UNIREP tests. Corresponding estimated covariance matrices were calculated, as well as lower and upper bounds for power using the methods presented in section 4.5. Approximate confidence interval coverage was defined as the proportion of the $500{,}000$ simulated bound realizations that successfully covered the observed mean power values for each condition described above. Only coverage of observed mean power values, and not predicted, were tabulated. The accuracy of the predicted power values, with respect to the observed, made it essentially redundant to consider both. Both one- and two-sided confidence intervals were evaluated with target coverages of 90% and 95%.

In practical biomedical research, low power values are of little concern. Rarely will one have a power targeted below 0.70. Therefore, only the results for target power values of 0.80 will be presented and discussed. Power confidence interval coverage converged to the target coverage as sample size increased. Only the worst case results for two-sided 95% confidence intervals are presented here. The worst cases occurred with the smallest sample size for the target study, for a variety of population sphericity values and estimated population powers.

In Table 4.7, the observed mean population powers are presented for the four UNIREP tests for the population sphericity values and ranks of $\boldsymbol{X}$ considered for target rank-adjusted Huynh-Feldt power of 0.80 and sample size of 16. In general, as the population sphericity increased and rank of $\boldsymbol{X}$ increased, the observed mean power values for the Box conservative, the Geisser-Greenhouse and the rank-adjusted Huynh-Feldt tests decreased. Only the Box conservative had severely biased power values as the population sphericity increased.

In Table 4.8, the proportion of simulations in which the estimated confidence interval successfully covered the observed mean population power values for each test is shown. The results are based on using an estimating study with sample size, N_{est}, of 16 and rank of $\boldsymbol{X}$, rank($\boldsymbol{X}_{\text{est}}$), of 4. In general, the approximate power confidence intervals nearly always reached the target 95% coverage for the Box conservative test. The coverage became more conservative as rank of $\boldsymbol{X}$ decreased. Similarly, the coverage became more conservative for the Geisser-Greenhouse and rank-adjusted Huynh-Feldt tests as rank of $\boldsymbol{X}$ decreased. The Geisser-Greenhouse and rank-adjusted Huynh-Feldt tests performed adequately in all cases except for the midrange population sphericity value, $\epsilon = 0.505$. The largest deviation from the target 95% estimated coverage was 13.6% and 16.0% for the Geisser-Greenhouse and Huynh-Feldt tests, respectively, which occurred for $\epsilon = 0.505$ and rank of $\boldsymbol{X}$ equal to 8. The approximate power confidence intervals for the uncorrected test reached the target coverage value for every case considered in which the uncorrected test would be used.

Although not presented here, in general, as sample size increased the conservative coverage values observed for the Box conservative and the uncorrected tests slowly converged to the target coverage value. This trend was observed for the conservative coverage values with the extreme population sphericity values for the Geisser-Greenhouse and the Huynh-Feldt tests as well. The same is true of the liberal coverage values observed for the midrange population sphericity values for the Geisser-Greenhouse and the Huynh-

Feldt tests. Similar results were obtained for the target 90% two-sided confidence interval coverage as well as the 95% and 90% one-sided confidence intervals coverage.

Table 4.7. Simulated Population Powers ($\times$ 100) for Target Power $=$ 80 with $N = 16$ and Rank$(X) = q$, Standard Error of Observed < 0.001.

	$\epsilon = 0.282$			$\epsilon = 0.505$		
q	**Box**	**GG**	**HF**	**Box**	**GG**	**HF**
2	77.9	81.1	81.7	56.1	77.8	80.9
4	76.3	79.7	80.5	51.0	76.2	80.2
8	75.3	78.7	79.9	45.5	73.6	79.6

	$\epsilon = 0.720$			$\epsilon = 1.00$			
	Box	**GG**	**HF**	**Box**	**GG**	**HF**	**UN**
2	45.7	76.0	80.5	39.9	74.8	79.0	79.9
4	35.5	74.0	80.1	25.5	72.4	78.7	80.1
8	26.7	69.5	79.5	13.8	65.5	77.5	80.0

Table 4.8. Target 95% CI (Two-Sided) Estimated Coverage ($\times$ 100) of Simulated Population Powers for Target Power $=$ 80 with $N = 16$ and Rank$(X) = q$. Estimation Study: $N_{\text{est}} = 16$ and Rank$(X_{\text{est}}) = 4$, 95% Half Confidence Interval is 6.04×10^{-4}.

	$\epsilon = 0.282$			$\epsilon = 0.505$		
q	**Box**	**GG**	**HF**	**Box**	**GG**	**HF**
2	97.8	97.2	97.0	97.5	93.4	92.3
4	93.7	92.0	91.6	95.6	86.8	85.0
8	90.9	87.9	87.2	94.8	81.4	79.0

	$\epsilon = 0.720$			$\epsilon = 1.00$			
	Box	**GG**	**HF**	**Box**	**GG**	**HF**	**UN**
2	97.6	95.4	94.9	97.4	95.3	95.5	95.8
4	97.5	93.6	92.9	97.6	96.8	97.0	97.4
8	96.9	90.6	89.8	97.0	96.1	96.9	97.4

Holding the target sample size constant, additional simulations were performed in an attempt to better understand the role of the estimating study in the calculation of approximate power confidence intervals. For a target study with sample size of 48 and rank of $\boldsymbol{X}$ equal to 2, 4, 8 and 16, observed mean power values for each UNIREP test were

tabulated for 500,000 replications for each of the conditions described above. Once again, only the cases with target rank-adjusted Huynh-Feldt power of 0.80 will be presented and discussed. In Table 4.9, the observed mean population powers are presented for the four UNIREP tests, the population sphericity values and ranks of $\boldsymbol{X}$ considered for sample size of 48. The information presented in this table is similar to that in Table 4.7.

The estimated coverages of these tabulated observed mean power values for each test are shown in Tables 4.10-4.13 for population sphericity values of 0.282, 0.505, 0.720 and 1.00, respectively. Approximate confidence intervals were simulated for 5,000 replications per condition (standard error of observed coverage less than or equal to 0.003). The estimating studies use sample sizes, N_{est}, of 16, 32 and 48, and ranks of $\boldsymbol{X}_{\text{est}}$ of 2, 4 and 8.

In general, for population sphericity values of 0.282 and 0.505, the approximate power confidence interval coverage for the Box conservative test converged to the target coverage value as rank of $\boldsymbol{X}_{\text{est}}$ increased, and thus ν_{est} decreased. Coverage decreased as rank of $\boldsymbol{X}$ from the target study increased. For larger rank of $\boldsymbol{X}$, the approximate power confidence interval coverage fell short of the target coverage in several instances. No clear trend was apparent as N_{est} increased. The Box conservative test would not be used for larger population sphericity values. However, the realization that the target coverage was reached in nearly every case considered for the larger population sphericity values is worth mentioning.

The approximate power confidence interval coverages for both the Geisser-Greenhouse and rank-adjusted Huynh-Feldt tests seem to have converged to the target coverage value as rank of $\boldsymbol{X}_{\text{est}}$ increased, and thus ν_{est} decreased, except in cases of sphericity. Such cases have little practical importance since exact results are available if sphericity is valid. Coverage decreased as rank of $\boldsymbol{X}$ from the target study increased. As observed in previous simulations, the approximate power confidence interval coverages for both the Geisser-Greenhouse and rank-adjusted Huynh-Feldt tests fell short of the target

coverage to varying degrees in nearly every case considered for midrange population sphericity values. This outcome was also observed for larger rank of $\boldsymbol{X}$ from the target study for population sphericity of 0.282. The approximate power confidence interval coverage for the uncorrected tests reached the target coverage value in every case except for large ν_{est} and small rank of $\boldsymbol{X}$ from the target study. The approximate power confidence interval coverage increased as the ranks of $\boldsymbol{X}$ for both the target and estimating studies increased and as N_{est} decreased.

The slow convergence of the approximate power confidence interval coverage to the target coverage may be due, in part, to use of $\widetilde{\epsilon}_n$ and $\widetilde{\epsilon}_r$ in the approximate power confidence interval equation. These estimators of the sphericity parameter are ratios of unbiased estimators for the nonnull and null cases, respectively. The variances of these estimators are much larger than the variances for $\widehat{\epsilon}_n$ and $\widehat{\epsilon}_d$. The larger variances may account for the slow convergence to the population power as the target and estimating study sample sizes and degrees of freedom increase. Further simulations may be needed to confirm this reasoning.

Table 4.9. Simulated Population Powers ($\times$ 100) for Target Power $=$ 80 with $N = 48$ and Rank$(X) = q$, Standard Error of Observed $<$ 0.001.

	$\epsilon = 0.282$			$\epsilon = 0.505$		
q	**Box**	**GG**	**HF**	**Box**	**GG**	**HF**
2	80.3	84.3	84.5	58.6	80.2	81.3
4	77.3	81.2	81.5	55.2	79.4	80.5
8	76.6	80.0	80.3	54.4	78.7	79.9
16	76.2	79.6	79.9	52.2	78.0	79.5

	$\epsilon = 0.720$			$\epsilon = 1.00$			
	Box	**GG**	**HF**	**Box**	**GG**	**HF**	**UN**
2	50.0	79.2	80.6	45.5	78.5	79.7	80.0
4	42.7	78.7	80.2	34.6	78.1	79.6	80.0
8	40.2	78.1	79.9	28.0	77.8	79.7	80.1
16	35.9	77.5	79.9	22.1	77.0	79.5	80.1

Table 4.10. Target 95% CI (Two-Sided) Estimated Coverage (× 100) of Simulated Population Powers for Population $\epsilon = 0.282$ for Target Power = 80 with $N = 48$ and Rank$(X) = q$. Estimation Study: $N_{\text{est}} \in (16, 32, 48)$ and Rank$(X_{\text{est}}) \in (2, 4, 8)$, 95% Half Confidence Interval is 6.04×10^{-3}.

		Box Coverage			GG Coverage			HF Coverage		
		Rank(X_{est})			Rank(X_{est})			Rank(X_{est})		
N_{est}	q	2	4	8	2	4	8	2	4	8
16	2	97.0	96.6	96.5	96.0	95.9	95.5	95.9	95.8	95.9
	4	92.5	92.7	92.7	91.0	90.9	91.3	90.5	90.7	90.5
	8	87.8	87.9	88.6	86.0	86.1	86.8	85.2	85.1	85.9
	16	85.9	86.6	86.9	84.6	84.6	85.4	84.4	84.4	85.3
32	2	96.8	96.8	96.6	95.1	95.3	95.7	95.7	95.7	96.1
	4	92.1	91.6	92.1	90.0	89.6	90.0	89.5	89.5	90.3
	8	86.7	87.1	87.4	85.1	85.1	85.1	84.7	85.1	85.5
	16	87.2	87.0	86.8	83.9	83.6	83.8	83.1	82.9	83.2
64	2	96.8	96.8	96.9	95.3	95.1	95.3	95.6	95.5	95.4
	4	91.4	91.5	91.6	89.9	90.2	90.0	90.0	89.8	89.9
	8	87.9	87.6	87.5	84.6	84.8	84.9	83.7	84.2	84.1
	16	86.9	86.7	86.6	83.2	83.6	83.3	83.7	83.9	83.6

Table 4.11. Target 95% CI (Two-Sided) Estimated Coverage (× 100) of Simulated Population Powers for Population $\epsilon = 0.505$ for Target Power = 80 with $N = 48$ and Rank$(X) = q$. Estimation Study: $N_{\text{est}} \in (16, 32, 48)$ and Rank$(X_{\text{est}}) \in (2, 4, 8)$, 95% Half Confidence Interval is 6.04×10^{-3}.

		Box Coverage			GG Coverage			HF Coverage		
		Rank(X_{est})			Rank(X_{est})			Rank(X_{est})		
N_{est}	q	2	4	8	2	4	8	2	4	8
16	2	97.5	97.2	97.4	94.1	94.3	94.9	92.2	92.1	93.6
	4	94.8	94.8	95.3	87.3	87.5	88.9	85.9	86.2	87.4
	8	92.6	92.7	93.4	83.2	83.4	86.0	81.4	82.0	84.3
	16	92.0	92.3	93.7	82.3	82.5	85.9	80.5	80.7	83.2
32	2	97.3	97.3	97.3	93.3	93.3	93.4	92.4	92.5	92.6
	4	93.8	94.1	94.3	85.7	85.2	85.8	85.0	84.8	84.5
	8	91.6	91.8	91.4	81.3	81.5	82.4	79.5	80.0	80.6
	16	91.5	91.7	91.7	79.4	78.9	80.0	79.2	79.1	79.5
64	2	97.2	97.2	97.4	93.6	93.7	93.5	92.6	92.5	92.8
	4	94.4	94.6	94.8	84.5	85.0	84.7	84.4	85.0	85.2
	8	91.7	91.5	91.8	79.6	80.1	80.4	78.9	79.2	79.6
	16	90.9	90.9	91.0	78.5	78.4	78.7	78.9	78.4	78.7

Table 4.12. Target 95% CI (Two-Sided) Estimated Coverage (× 100) of Simulated Population Powers for Population $\epsilon = 0.720$ for Target Power = 80 with $N = 48$ and Rank$(X) = q$. Estimation Study: $N_{\text{est}} \in (16, 32, 48)$ and Rank$(X_{\text{est}}) \in (2, 4, 8)$, 95% Half Confidence Interval is 6.04×10^{-3}.

		Box Coverage			GG Coverage			HF Coverage		
		Rank(X_{est})			Rank(X_{est})			Rank(X_{est})		
N_{est}	q	2	4	8	2	4	8	2	4	8
16	2	97.5	97.3	97.7	94.5	94.4	94.8	95.1	95.4	95.8
	4	96.5	96.9	97.8	93.4	94.0	93.6	93.6	93.9	95.1
	8	96.3	96.6	97.1	90.7	91.1	91.6	91.0	91.5	93.1
	16	96.2	96.3	97.2	89.7	89.9	90.7	89.9	90.2	91.8
32	2	96.8	96.9	97.3	93.5	93.6	93.5	94.5	94.6	94.8
	4	96.4	96.3	96.4	92.2	92.4	92.2	92.0	91.8	91.9
	8	95.4	95.4	95.5	89.4	89.4	89.6	88.9	88.9	89.2
	16	95.0	94.8	95.2	88.9	88.3	88.8	87.9	88.6	88.5
64	2	96.3	96.6	96.8	93.3	93.2	93.3	93.1	92.8	93.0
	4	95.8	95.8	95.9	91.8	92.1	92.0	91.0	91.1	90.6
	8	95.3	95.5	95.1	88.2	88.5	88.5	88.2	87.6	88.1
	16	94.6	94.5	94.7	87.4	87.5	87.5	86.5	86.6	86.7

Table 4.13. Target 95% CI (Two-Sided) Estimated Coverage (× 100) of Simulated Population Powers for Population $\epsilon = 1.00$ for Target Power = 80 with $N = 48$ and Rank$(X) = q$. Estimation Study: $N_{\text{est}} \in (16, 32, 48)$ and Rank$(X_{\text{est}}) \in (2, 4, 8)$, 95% Half Confidence Interval is 6.04×10^{-3}.

		Box Coverage			GG Coverage			HF Coverage			UN Coverage		
		Rank(X_{est})			Rank(X_{est})			Rank(X_{est})			Rank(X_{est})		
N_{est}	q	2	4	8	2	4	8	2	4	8	2	4	8
16	2	97.3	97.1	98.0	93.4	93.1	92.9	95.3	95.2	96.0	95.5	95.6	96.5
	4	97.7	97.8	98.5	94.7	94.6	94.6	97.4	97.5	97.9	97.2	97.6	98.3
	8	97.3	97.6	98.2	94.9	94.7	94.5	96.9	96.9	97.5	97.3	97.4	98.3
	16	97.5	97.6	98.5	94.8	94.9	94.5	96.7	96.8	97.2	97.3	97.4	98.3
32	2	96.4	96.5	96.5	93.6	93.8	93.6	93.5	93.3	93.5	94.0	94.0	94.5
	4	96.4	96.5	96.6	95.9	95.6	95.8	96.0	96.3	96.7	95.8	96.1	96.5
	8	96.5	96.5	96.6	95.1	95.1	95.0	96.4	96.5	96.5	96.3	96.1	96.6
	16	96.4	96.5	96.7	95.7	95.9	96.2	95.7	95.9	96.2	96.8	96.7	96.7
64	2	95.6	95.9	95.9	92.7	92.8	93.1	93.2	93.3	93.3	92.8	92.7	92.8
	4	95.6	95.5	95.7	94.7	94.7	95.0	95.4	95.5	95.7	95.8	95.7	96.1
	8	95.9	96.0	96.1	95.3	95.6	95.4	95.7	95.7	95.5	95.6	95.8	95.8
	16	95.8	96.0	96.0	94.6	94.7	95.1	95.6	95.3	95.5	95.6	95.8	96.2

4.7 Alternative Approximations Considered for Estimated Covariance

In attempts to develop better confidence bound estimates for UNIREP power, additional experimental fittings of a variety of different distributions to $\widetilde{\lambda}_{*1}$, or a function of it, have been performed. One attempt was to approximate the distribution of $\widetilde{\lambda}_{*1}$ with an F. Fitting $\widetilde{\lambda}_{*1}$ to an F made sense for several reasons. First, an F distribution contains the correct support. All the components of $\widetilde{\lambda}_{*1}$ are positive suggesting that $0 \leq \widetilde{\lambda}_{*1}$. Second, $\widetilde{\lambda}_{*1}$ is a ratio of two variables that could be somewhat estimated by chi-squares.

Using the methods presented in Kim *et al.* (2006), the numerator of $\widetilde{\lambda}_{*1}$ was approximated with a weighted noncentral chi-square, while the denominator was approximated with a weighted central chi-square. Two concerns arose. First, the denominator is not necessarily a central quadratic. The $2\text{tr}(\boldsymbol{\Delta}/a)$ component makes the denominator more of a shifted central quadratic. Second, the Kim *et al.* (2006) result

requires that the components of the numerator and denominator be mutually independent. This requirement is not met. As a result, the approximated variance was much larger than the simulated cases, which led to a poor distributional fit.

Additional attempts to match only the numerator to a weighted noncentral chi-square or to a weighted central chi-square with the denominator a constant equal to $\mathrm{E}[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*) + 2\mathrm{tr}(\boldsymbol{\Delta}/a)]$ were performed. These attempts resulted in similarly inaccurate outcomes.

4.8 Conclusions

In practice, statisticians realize that a measure of uncertainty that can be associated with a parameter estimate is important. When the random parameter is power, confidence intervals that account for the uncertainty provide a method to state that a study has power of at least "P" to detect an effect, with a specified confidence. For an estimated variance and fixed means, methods to provide exact power confidence intervals exist in the univariate setting. In this paper, approximate power confidence intervals have been proposed and evaluated in the UNIREP setting for an estimated covariance and fixed means. The methods have been evaluated for a large range of conditions, and have been shown to provide reasonably accurate coverage for power for all four UNIREP tests.

Even for small sample sizes, the proposed power confidence intervals attain very accurate coverage probabilities for the Box conservative and uncorrected tests in all cases. This result is also true for the extreme population sphericity values for the Geisser-Greenhouse and Huynh-Feldt tests. For midrange population sphericity values, the coverage probabilities of the approximate power confidence intervals for the latter two tests often fell short of the various target coverage values considered. Still, these results are quite good, considering the small sample sizes being considered. Coverage probabilities did improve as sample size increased, as demonstrated through additional simulations (not all presented here). The approximate confidence intervals performed better for higher target

power values than for lower. This realization makes the approximate confidence intervals more useful in practical research conditions. One-sided confidence intervals are recommended to provide tighter, more informative confidence bounds, as compared to the two-sided confidence intervals.

The techniques also provide the means to plot power confidence regions around an estimated power curve as demonstrated in Figures 4.1-4.4. The resulting method of displaying power analysis results have been extremely well received by researchers who have seen it.

Many factors play a role in the computation of these approximate power confidence intervals. In general, the approximate power confidence interval coverage converged to the target coverage value as the target and estimating study sample sizes increased. As rank of $\boldsymbol{X}$ for the target and estimating studies increased, the coverage probabilities seem to have decreased. The coverage reduction was particularly noticeable for the Geisser-Greenhouse and Huynh-Feldt tests for midrange population sphericity values. These comments are merely generalities, however. The interplay of sample size, rank of $\boldsymbol{X}$ and the degrees of freedom for the estimating study confounded the trends in some cases.

The estimated sphericity parameters involved in the calculation of the approximate power confidence intervals may also have confounded the general trends by having a stronger influence over the coverage probabilities under certain conditions. The approximately unbiased estimators for both the nonnull and null cases have larger variances than their corresponding MLEs. The approximately unbiased estimators are integrated into the approximate power confidence interval function at several points. Their larger variances almost certainly play a role in the slow convergence to the population power, and thus target coverage.

Despite the many parts of the approximation, simulation results provide evidence that the methods allow one to calculate reasonably accurate and useful power confidence

intervals for all four UNIREP tests. One must remember that a lot of guess work is often involved in power analyses. With that in mind, it is reasoned that the approximations perform well enough for nearly all practical uses, even for smaller sample sizes, and should be used in future study designs employing UNIREP techniques.

Chapter 5
Conclusions and Recommendations for Future Research

5.1 Conclusions

UNIREP techniques make up a special case of the more broad area of statistical modeling called mixed models. Due, in part, to their good inference and power techniques, UNIREP analyses should be used whenever possible. Here, three areas related to power for UNIREP tests have been examined and improved upon. Although the new methods introduced here apply to a wide variety of studies, such as experimental or controlled laboratory research, the driving motivation and application has been imaging research. Imaging research often generates the type of complete data that can be handled with UNIREP procedures. Also, such research often involves small sample sizes, which makes UNIREP (or MULTIREP) techniques much more desirable than mixed models.

The Huynh-Feldt sphericity estimator was developed in an attempt to correct for biases found with the Geisser-Greenhouse estimator. Huynh and Feldt (1976) claimed that their estimator was a ratio of unbiased estimators. They further asserted that their estimator was less biased and less dependent on large sample sizes than the Geisser-Greenhouse estimator when the population covariance deviated only moderately from sphericity. In Chapter 2, their claims are shown to be true only for the special case of rank of $\boldsymbol{X}$ equal to 1.

In Chapter 2, a rank-adjusted Huynh-Feldt sphericity estimator was introduced and evaluated for a wide range of conditions. When rank of $\boldsymbol{X}$ was greater than 1, the rank-adjusted estimator was shown to better estimate the population sphericity than the Huynh-Feldt estimator. This outcome was particularly true for larger rank of $\boldsymbol{X}$. The biased Huynh-Feldt estimator yields a biased Huynh-Feldt test. Furthermore, for any rank of $\boldsymbol{X}$,

the rank-adjusted estimator was shown to be less biased and less dependent on large sample sizes than the Geisser-Greenhouse estimator when the population covariance deviated only moderately from sphericity. The rank-adjusted estimator is a ratio of two unbiased estimators for any rank of $\boldsymbol{X}$, and reduces to the Huynh-Feldt estimator when rank of $\boldsymbol{X}$ is equal to 1. In this sense, the rank-adjusted estimator is more theoretically in line with the goals originally set forth by Huynh and Feldt (1976).

Muller *et al.* (2007) introduced approximate power calculations for all four UNIREP tests, which were accurate and easy to use. Their power approximation for the Huynh-Feldt test incorporates the Huynh-Feldt sphericity estimator. In Chapter 3, the work begun in Chapter 2 was extended by introducing a power approximation for the rank-adjusted test, which uses the rank-adjusted sphericity estimator. The accuracy of the rank-adjusted power approximation was evaluated for a wide range of conditions. For practical research situations, the rank-adjusted power approximation was shown to perform as well as and, in most cases, better than the Huynh-Feldt power approximation. Furthermore, the Huynh-Feldt power approximation was shown to yield artificially inflated power values at a cost of inflated test size when rank of $\boldsymbol{X}$ was greater than 1. In some cases, the Huynh-Feldt test size was observed to be greater than double the target test size. Meanwhile, the rank-adjusted test controlled test size adequately. Based on the results presented in Chapters 2 and 3, use of the rank-adjusted sphericity estimator, test and power approximation are recommended over the corresponding Huynh-Feldt methods.

Accurate power analysis is essential when designing a study. An accurate power analysis allows researchers the ability to focus study hypotheses, clarify the analysis plans and enhance study design efficiency. Power is computed assuming known values of distributional parameters. Rarely is the variance actually known in these computations. Rather, the variance is often estimated from previous studies. When the variance is

estimated, power becomes a random variable. Providing confidence intervals that account for the uncertainty in these random power values would be useful in any study design.

For estimated variance and fixed means, exact power confidence intervals for univariate analyses have been presented by Taylor and Muller (1995). In Chapter 4, the methods of Taylor and Muller (1995) were extended to provide reasonably accurate, approximate confidence intervals for UNIREP power, in the case of an estimated covariance and fixed means. The approximate confidence intervals performed well in most cases considered, even for small sample sizes. The approximate confidence intervals performed better for higher target power values than for lower. This realization makes the approximate confidence intervals more useful in practical research conditions. For midrange population sphericity values, the coverage probabilities of the approximate confidence intervals for the Geisser-Greenhouse and Huynh-Feldt tests often fell short of the target coverage. However, coverage probabilities for all four UNIREP tests seemed to converge to the target coverage value as sample size for both the target and estimating studies increased. As rank of $\boldsymbol{X}$ for the target and estimating studies increased, the coverage probabilities seem to have decreased. One-sided confidence intervals are recommended to provide tighter, more informative confidence bounds, as compared to the two-sided confidence intervals.

The methods presented in Chapter 4 offer more than merely approximate confidence intervals for a single power value at a time in the UNIREP setting. The methods may be extended to allow for calculation and graphical representation of accurate confidence *regions* for the entire power curve. The resulting method of displaying power analysis results have been extremely well received by researchers who have seen it.

5.2 Recommendations for Future Research

While the methods presented here improve upon several areas for power for UNIREP tests, they also introduce new questions that may be examined in future research. The

methods discussed focus solely on UNIREP procedures. The reason for such a focus was to take advantage of the good power methods that accompany UNIREP procedures, especially for small sample sizes. Power techniques for UNIREP have been well tested and documented. This is not as true for the general mixed model. Much of imaging research and experimental or controlled laboratory studies do not require the analysis qualities that are associated with mixed model procedures. However, mixed models are called for in many research situations, and easy to use software is readily available.

The mixed model has several nice statistical features, such as no requirement for balanced data, the ability to explicitly model and analyze the between- and within-subject variation, and the capability of handling missing data without eliminating all values for a particular subject. Furthermore, mixed models allow researchers to specify the type of covariance structure desired. In this respect, they are very convenient to use.

Proper use of UNIREP procedures assumes complete data. UNIREP methods were generalized by Catellier and Muller (2000) to allow for missing data. This progression seems to be a natural one. The expectation is that the research introduced here will lay the groundwork for future researchers to ultimately extend these methods to fit with the general mixed model. Accurate power and power confidence intervals in research studies with missing, mistimed or unbalanced data would benefit researchers from every field of study.

Several specific challenges must be overcome in order to apply the methods described here to studies with missing data. The challenges include, but are not limited to, an examination of the effects of percent of missing data in a research study, as well as types of missing data. The types of missing data may include missing at random (MAR) or missing completely at random (MCAR). The challenge of estimating the covariance matrix with missing data may also be a potential source of difficulty. A study with mistimed data may be thought of as one with extreme amounts of missing data. It seems doubtful that the

methods described here would work well in such a setting because UNIREP tests implicitly require estimating a complete unstructured covariance matrix.

In section 2.4, the proportion of observed truncated estimates of both the Huynh-Feldt and rank-adjusted sphericity estimators were examined for the simulation cases considered. The examination provides only the first step. Evaluations of the entire distributions of the estimators are recommended. Information about the distributions of the Huynh-Feldt, rank-adjusted and Geisser-Greenhouse sphericity estimates would provide a better understanding of sphericity estimates and their effects on test size and power.

Many factors play a role in the computation of the approximate power confidence intervals for UNIREP tests introduced in Chapter 4. The interplay of sample size, rank of $\boldsymbol{X}$ and the degrees of freedom for the estimating and target studies confounded the trends of confidence interval coverage in some cases. The estimated sphericity parameters involved in the calculation of the approximate power confidence intervals may also have confounded the general trends by having a stronger influence over the coverage probabilities under certain conditions.

The approximate power confidence intervals performed well in most cases considered. However, some of the simulated results leave room for improvement. The confidence interval coverage for both the Geisser-Greenhouse and Huynh-Feldt tests for midrange population sphericity values consistently fell below target coverage values. Also, the observed coverage probabilities for all UNIREP tests seemed to decrease as rank of $\boldsymbol{X}$ increased for both the estimating and target studies. This decrease in coverage probabilities resulted in coverage values falling below the target coverage values. A better understanding of these trends require further examination of the accuracy of these approximate power confidence intervals for various conditions.

The convergence of the approximate power confidence interval coverage to target coverage values were surprisingly slow as sample size for the target and estimating studies

increased. The approximately unbiased estimators for both the nonnull and null cases are integrated into the approximate power confidence interval function at several points. The approximately unbiased estimators for both the nonnull and null cases have larger variances than their corresponding MLEs. Their larger variances almost certainly play a role in the slow convergence to the population power, and thus target coverage. Further examination of these estimators and their effects are needed. Perhaps appropriate correction factors may be developed and applied under certain conditions.

Finally, the realization that the variance component used in the approximate power confidence intervals for the UNIREP tests is approximated to fit a chi-square distribution, rather than fit exactly, allows for the possibility of a better approximation. In section 4.7, additional distributional approximations were discussed and discounted due to their poor results. The approximate power confidence intervals performed well in most cases considered. However, perhaps a better approximate distributional fit would allow for more accurate results. This task is left for future research.

Appendix A: Chapter 2 Proofs

Lemma 2.1 Unbiased estimators for $\tau_1 = \text{tr}^2(\boldsymbol{\Sigma}_*)$ and $\tau_2 = \text{tr}(\boldsymbol{\Sigma}_*^2)$ are

$$\begin{aligned}\widehat{\tau}_1 &= [t_1 - 2(\nu_e+1)^{-1}t_2]\nu_e^{-2}\{1 - 2[\nu_e(\nu_e+1)]^{-1}\}^{-1} \\ &= [\widetilde{\tau}_1 - 2(\nu_e+1)^{-1}\widetilde{\tau}_2]\{1 - 2[\nu_e(\nu_e+1)]^{-1}\}^{-1}\end{aligned} \tag{A.1}$$

$$\begin{aligned}\widehat{\tau}_2 &= (\nu_e t_2 - t_1)\{\nu_e[\nu_e(\nu_e+1) - 2]\}^{-1} \\ &= \left(\nu_e^2\widetilde{\tau}_2 - \nu_e\widetilde{\tau}_1\right)[\nu_e(\nu_e+1) - 2]^{-1}\,.\end{aligned} \tag{A.2}$$

Proof. Suppose $t_1 = \text{tr}^2(\boldsymbol{S}_e) = \nu_e^2\widetilde{\tau}_1 = \nu_e^2\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)$ and $t_2 = \text{tr}(\boldsymbol{S}_e^2) = \nu_e^2\widetilde{\tau}_2 = \nu_e^2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)$.

Muller *et al.* (2007, Appendix A) showed that

$$\begin{aligned}\text{E}(t_1|\nu_e \geq b) = \text{E}\left[\text{tr}^2(\boldsymbol{S}_e)|\nu_e \geq b\right] &= 2\nu_e\sum\nolimits_{k=1}^{b}\lambda_k^2 + \nu_e^2(\sum\nolimits_{k=1}^{b}\lambda_k)^2 \\ &= 2\nu_e\boldsymbol{\lambda}'\boldsymbol{\lambda} + \nu_e^2(\mathbf{1}_b'\boldsymbol{\lambda})^2 \\ &= 2\nu_e\tau_2 + \nu_e^2\tau_1 \\ &= 2\nu_e\text{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e^2\text{tr}^2(\boldsymbol{\Sigma}_*)\end{aligned}$$

and

$$\begin{aligned}\text{E}(t_2|\nu_e \geq b) = \text{E}\left[\text{tr}(\boldsymbol{S}_e^2)|\nu_e \geq b\right] &= \nu_e(\nu_e+2)\sum_{k_1=1}^{b}\lambda_{k_1}^2 + 2\nu_e\sum_{k_1=2}^{b}\sum_{k_2=1}^{k_1-1}\lambda_{k_1}\lambda_{k_2} \\ &= \nu_e(\nu_e+2)\boldsymbol{\lambda}'\boldsymbol{\lambda} + \nu_e(\mathbf{1}_b'\boldsymbol{\lambda}\boldsymbol{\lambda}'\mathbf{1}_b - \boldsymbol{\lambda}'\boldsymbol{\lambda}) \\ &= \nu_e(\nu_e+1)\boldsymbol{\lambda}'\boldsymbol{\lambda} + \nu_e(\mathbf{1}_b'\boldsymbol{\lambda})^2 \\ &= \nu_e(\nu_e+1)\tau_2 + \nu_e\tau_1 \\ &= \nu_e(\nu_e+1)\text{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e\text{tr}^2(\boldsymbol{\Sigma}_*)\,.\end{aligned}$$

Thus, $\widehat{\tau}_1$ and $\widehat{\tau}_2$ are unbiased estimators of τ_1 and τ_2, as demonstrated below:

$$
\begin{aligned}
\mathrm{E}(\widehat{\tau}_1) &= [\mathrm{E}(t_1) - 2(\nu_e+1)^{-1}\mathrm{E}(t_2)]\nu_e^{-2}\{1-2[\nu_e(\nu_e+1)]^{-1}\}^{-1} \\
&= \left\{\left[2\nu_e \mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e^2\mathrm{tr}^2(\boldsymbol{\Sigma}_*)\right] - 2(\nu_e+1)^{-1}\left[\nu_e(\nu_e+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e\mathrm{tr}^2(\boldsymbol{\Sigma}_*)\right]\right\} \times \\
&\qquad \nu_e^{-2}\left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}^{-1} \\
&= \left\{\frac{[2\nu_e\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e^2\mathrm{tr}^2(\boldsymbol{\Sigma}_*)]}{\nu_e^2} - \frac{[2\nu_e(\nu_e+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + 2\nu_e\mathrm{tr}^2(\boldsymbol{\Sigma}_*)]}{\nu_e^2(\nu_e+1)}\right\} \times \\
&\qquad \left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}^{-1} \\
&= \left\{\frac{\nu_e(\nu_e+1)[2\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e\mathrm{tr}^2(\boldsymbol{\Sigma}_*)]}{\nu_e^2(\nu_e+1)} - \frac{[2\nu_e(\nu_e+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + 2\nu_e\mathrm{tr}^2(\boldsymbol{\Sigma}_*)]}{\nu_e^2(\nu_e+1)}\right\} \times \\
&\qquad \left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}^{-1} \\
&= \left\{\frac{\nu_e^2(\nu_e+1)\mathrm{tr}^2(\boldsymbol{\Sigma}_*) - 2\nu_e\mathrm{tr}^2(\boldsymbol{\Sigma}_*)}{\nu_e^2(\nu_e+1)}\right\}\left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}^{-1} \\
&= \mathrm{tr}^2(\boldsymbol{\Sigma}_*)\left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}\left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}^{-1} \\
&= \mathrm{tr}^2(\boldsymbol{\Sigma}_*) \\
&= \tau_1
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(\widehat{\tau}_2) &= [\nu_e\mathrm{E}(t_2) - \mathrm{E}(t_1)]\{\nu_e[\nu_e(\nu_e+1)-2]\}^{-1} \\
&= \left\{\nu_e\left[\nu_e(\nu_e+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e\mathrm{tr}^2(\boldsymbol{\Sigma}_*)\right] - \left[2\nu_e\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e^2\mathrm{tr}^2(\boldsymbol{\Sigma}_*)\right]\right\} \times \\
&\qquad \{\nu_e[\nu_e(\nu_e+1)-2]\}^{-1} \\
&= \left\{\nu_e\nu_e(\nu_e+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e\mathrm{tr}^2(\boldsymbol{\Sigma}_*) - 2\nu_e\mathrm{tr}(\boldsymbol{\Sigma}_*^2) - \nu_e^2\mathrm{tr}^2(\boldsymbol{\Sigma}_*)\right\} \times \\
&\qquad \{\nu_e[\nu_e(\nu_e+1)-2]\}^{-1} \\
&= \left\{\nu_e\nu_e(\nu_e+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) - 2\nu_e\mathrm{tr}(\boldsymbol{\Sigma}_*^2)\right\}\{\nu_e[\nu_e(\nu_e+1)-2]\}^{-1} \\
&= \mathrm{tr}(\boldsymbol{\Sigma}_*^2)\{\nu_e[\nu_e(\nu_e+1)-2]\}\{\nu_e[\nu_e(\nu_e+1)-2]\}^{-1} \\
&= \mathrm{tr}(\boldsymbol{\Sigma}_*^2) \\
&= \tau_2
\end{aligned}
$$

□

Lemma 2.2 A ratio estimating ϵ in terms of correlated, but unbiased, estimators is

$$
\begin{aligned}
\widetilde{\epsilon}_r &= \frac{\widehat{\tau}_1}{b\widehat{\tau}_2} \\
&= \frac{(\nu_e+1)b\widehat{\epsilon} - 2}{b(\nu_e - b\widehat{\epsilon})}\,.
\end{aligned}
\tag{A.3}
$$

Proof. Unbiased estimators for $\mathrm{tr}^2(\boldsymbol{\Sigma}_*)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_*^2)$ are

$$\widehat{\tau}_1 = [\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2(\nu_e+1)^{-1}\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)]\{1-2[\nu_e(\nu_e+1)]^{-1}\}^{-1}$$

$$\widehat{\tau}_2 = \left(\nu_e^2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)\right)[\nu_e(\nu_e+1)-2]^{-1}\,,$$

as introduced in Lemma 2.1. Thus,

$$\begin{aligned}
\widetilde{\epsilon}_r &= \frac{\widehat{\tau}_1}{b\widehat{\tau}_2} \\
&= \frac{[t_1 - 2(\nu_e+1)^{-1}t_2]\nu_e^{-2}\left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}^{-1}}{b(\nu_e t_2 - t_1)\{\nu_e[\nu_e(\nu_e+1)-2]\}^{-1}} \\
&= \frac{[t_1/t_2 - 2(\nu_e+1)^{-1}]\nu_e^{-2}\left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}^{-1}}{b(\nu_e - t_1/t_2)\{\nu_e[\nu_e(\nu_e+1)-2]\}^{-1}} \\
&= \frac{[t_1/t_2 - 2(\nu_e+1)^{-1}]\{\nu_e[\nu_e(\nu_e+1)-2]\}}{b(\nu_e - t_1/t_2)\nu_e^2\left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}} \\
&= \frac{[t_1/t_2 - 2(\nu_e+1)^{-1}][\nu_e(\nu_e+1)-2]}{b(\nu_e - t_1/t_2)\nu_e\left\{1-2[\nu_e(\nu_e+1)]^{-1}\right\}} \\
&= \frac{[(\nu_e+1)t_1/t_2 - 2][\nu_e(\nu_e+1)-2]}{b(\nu_e - t_1/t_2)\nu_e\{(\nu_e+1)-2\nu_e^{-1}\}} \\
&= \frac{[(\nu_e+1)t_1/t_2 - 2]}{b(\nu_e - t_1/t_2)} \\
&= \frac{(\nu_e+1)b\widehat{\epsilon}-2}{b(\nu_e - b\widehat{\epsilon})}
\end{aligned}$$

□

Corollary 2.3 If $\text{rank}(\boldsymbol{X}) = 1$, then $\widetilde{\epsilon}_r = \widetilde{\epsilon}_{HF}$, the estimator proposed by Huynh and Feldt (1976).

Proof. Here, $\text{rank}(\boldsymbol{X}) = 1$, which suggest $\nu_e = N - 1$. Then,

$$
\begin{aligned}
\widetilde{\epsilon}_r &= \frac{(\nu_e + 1)b\widehat{\epsilon} - 2}{b(\nu_e - b\widehat{\epsilon})} \\
&= \frac{(N - 1 + 1)b\widehat{\epsilon} - 2}{b(N - 1 - b\widehat{\epsilon})} \\
&= \frac{Nb\widehat{\epsilon} - 2}{b(N - 1 - b\widehat{\epsilon})} \\
&= \frac{Nb\widehat{\epsilon} - 2}{b(\nu_e - b\widehat{\epsilon})} \\
&= \widetilde{\epsilon}_{HF} \, .
\end{aligned}
$$

□

Appendix B: Chapter 4 Proofs

Lemma 4.1 Defining

$$S_{t1} = \sum_{k=1}^{b} \lambda_k \ , \quad S_{t2} = \sum_{k=1}^{b} \lambda_k^2 \ , \quad S_{t3} = \sum_{k=1}^{b} \lambda_k \omega_{*kk} \ , \quad S_{t4} = \sum_{k=1}^{b} \lambda_k^2 \omega_{*kk} \tag{B.1}$$

implies $\{S_{t1}, S_{t2}, S_{t3}, S_{t4}\} = \{\text{tr}(\boldsymbol{\Sigma}_*), \text{tr}(\boldsymbol{\Sigma}_*^2), \text{tr}(\boldsymbol{\Delta}), \text{tr}(\boldsymbol{\Sigma}_*\boldsymbol{\Delta})\}$.

Proof. Let $\omega_{*kk} = \boldsymbol{v}_k'\boldsymbol{\Delta}\boldsymbol{v}_k/\lambda_k$ be the diagonal elements of the scaled noncentrality, $\boldsymbol{\Omega}_* = \boldsymbol{\Upsilon}'\boldsymbol{\Delta}\boldsymbol{\Upsilon}\text{Dg}(\boldsymbol{\lambda})^{-1} = \boldsymbol{\Delta}_*\text{Dg}(\boldsymbol{\lambda})^{-1}$. Define $\boldsymbol{\Upsilon}$, $(b \times b)$, as a matrix of the eigenvectors of $\boldsymbol{\Sigma}_* = \boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{\Upsilon}\text{Dg}(\boldsymbol{\lambda})\boldsymbol{\Upsilon}'$, such that $\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}' = \boldsymbol{\Upsilon}'\boldsymbol{\Upsilon} = \boldsymbol{I}_b$. Also let λ_i be the eigenvalues for $\boldsymbol{\Sigma}_*$. Then,

$$\begin{aligned} S_{t1} &= \sum_{k=1}^{b}\lambda_k = \sum_{k=1}^{b}\text{tr}(\lambda_k \boldsymbol{I}_b) = \sum_{k=1}^{b}\text{tr}(\lambda_k \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}') = \sum_{k=1}^{b}\text{tr}(\boldsymbol{\Upsilon}'\lambda_k\boldsymbol{\Upsilon}) \\ &= \text{tr}\left[\sum_{k=1}^{b}(\boldsymbol{v}_k'\lambda_k\boldsymbol{v}_k)\right] = \text{tr}[(\boldsymbol{\Upsilon}'Dg(\lambda_k)\boldsymbol{\Upsilon})] = \text{tr}(\boldsymbol{\Sigma}_*) \ . \end{aligned}$$

$$\begin{aligned} S_{t2} &= \sum_{k=1}^{b}\lambda_k^2 = \sum_{k=1}^{b}\text{tr}\big(\lambda_k^2 \boldsymbol{I}_b\big) = \sum_{k=1}^{b}\text{tr}\big(\lambda_k^2 \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}'\big) = \sum_{k=1}^{b}\text{tr}\big(\boldsymbol{\Upsilon}'\lambda_k^2\boldsymbol{\Upsilon}\big) \\ &= \text{tr}\left[\sum_{k=1}^{b}\big(\boldsymbol{v}_k'\lambda_k^2\boldsymbol{v}_k\big)\right] = \text{tr}\big[\big(\boldsymbol{\Upsilon}'Dg\big(\lambda_k^2\big)\boldsymbol{\Upsilon}\big)\big] = \text{tr}(\boldsymbol{\Sigma}_*^2) \ . \end{aligned}$$

$$\begin{aligned} S_{t3} &= \sum_{k=1}^{b}\lambda_k\boldsymbol{v}_k'\boldsymbol{\Delta}\boldsymbol{v}_k/\lambda_k = \sum_{k=1}^{b}\text{tr}(\boldsymbol{v}_k'\boldsymbol{\Delta}\boldsymbol{v}_k) = \sum_{k=1}^{b}\text{tr}(\boldsymbol{\Delta}\boldsymbol{v}_k\boldsymbol{v}_k') \\ &= \text{tr}\left[\sum_{k=1}^{b}(\boldsymbol{\Delta}\boldsymbol{v}_k\boldsymbol{v}_k')\right] = \text{tr}\left[\boldsymbol{\Delta}\sum_{k=1}^{b}(\boldsymbol{v}_k\boldsymbol{v}_k')\right] = \text{tr}(\boldsymbol{\Delta}\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}') = \text{tr}(\boldsymbol{\Delta}) \ . \end{aligned}$$

Also,

$$S_{t4} = \sum_{k=1}^{b} \lambda_k^2 \boldsymbol{v}_k' \boldsymbol{\Delta} \boldsymbol{v}_k / \lambda_k = \sum_{k=1}^{b} \text{tr}(\lambda_k \boldsymbol{v}_k' \boldsymbol{\Delta} \boldsymbol{v}_k) = \sum_{k=1}^{b} \text{tr}(\boldsymbol{\Delta} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k')$$
$$= \text{tr}\left[\sum_{k=1}^{b} (\boldsymbol{\Delta} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k')\right] = \text{tr}\left[\boldsymbol{\Delta} \sum_{k=1}^{b} (\lambda_k \boldsymbol{v}_k \boldsymbol{v}_k')\right] = \text{tr}(\boldsymbol{\Delta} \boldsymbol{\Sigma}_*) \; .$$

□

Lemma 4.2 The constant in the critical value of the UNIREP test statistic approximation introduced by Muller *et al.* (2007) equals 1,

$$\frac{\lambda_{*2}}{\lambda_{*1}} \frac{ab}{\nu_{*1}} \frac{\nu_{*2}}{b\nu_e} = 1 \; . \tag{B.2}$$

Thus,

$$\Pr\{T_u \leq t\} \approx \Pr\left\{ \frac{\lambda_{*1} y_{*1}/(ab)}{\lambda_{*2} y_{*2}/(b\nu_e)} \leq t \right\} = F_F(t; \nu_{*1}, \nu_{*2}, \omega_*) \; . \tag{B.3}$$

Proof. Here, $\{S_{t1}, S_{t2}, S_{t3}, S_{t4}\} = \{\text{tr}(\boldsymbol{\Sigma}_*), \text{tr}(\boldsymbol{\Sigma}_*^2), \text{tr}(\boldsymbol{\Delta}), \text{tr}(\boldsymbol{\Sigma}_* \boldsymbol{\Delta})\}$ as in Lemma 4.1. Furthermore,

$$\begin{aligned}
\lambda_{*1} &= \frac{(aS_{t2} + 2S_{t4})}{(aS_{t1} + 2S_{t3})} \\
\nu_{*1} &= aS_{t1}/\lambda_{*1} \\
\omega_* &= S_{t3}/\lambda_{*1} \\
\lambda_{*2} &= S_{t2}/S_{t1} \\
\nu_{*2} &= \nu_e S_{t1}^2/S_{t2} = \nu_e b\epsilon
\end{aligned}$$

as in equations. 54-58. Then,

$$\begin{aligned}
\frac{\lambda_{*2}}{\lambda_{*1}} \frac{ab}{\nu_{*1}} \frac{\nu_{*2}}{b\nu_e} &= \frac{S_{t2}/S_{t1}}{\lambda_{*1}} \frac{ab}{aS_{t1}/\lambda_{*1}} \frac{\nu_e S_{t1}^2/S_{t2}}{b\nu_e} \\
&= \frac{S_{t2}}{\lambda_{*1} S_{t1}} \frac{ab\lambda_{*1}}{aS_{t1}} \frac{\nu_e S_{t1}^2}{b\nu_e S_{t2}} \\
&= \frac{ab\nu_e \lambda_{*1} S_{t1}^2 S_{t2}}{ab\nu_e \lambda_{*1} S_{t1}^2 S_{t2}} \\
&= 1
\end{aligned}$$

If $T_u = [\text{tr}(\widehat{\boldsymbol{\Delta}})/a]/[\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)]$, then $\text{tr}(\widehat{\boldsymbol{\Delta}}) \approx \lambda_{*1} y_{*1}$ and $\text{tr}(\widehat{\boldsymbol{\Sigma}}_*) \approx \lambda_{*2} y_{*2}$ with $y_{*1} \sim \chi^2(\nu_{*1}, \omega_*)$ and $y_{*2} \sim \chi^2(\nu_{*2})$ as described in Muller *et al.* (2007). Then,

$$\begin{aligned}
\Pr\{T_u \le t\} &= \Pr\left\{ \frac{\text{tr}(\widehat{\boldsymbol{\Delta}})/a}{\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)} \le t \right\} \\
&\approx \Pr\left\{ \frac{\lambda_{*1} y_{*1}/(ab)}{\lambda_{*2} y_{*2}/(b\nu_e)} \le t \right\} \\
&= \Pr\left\{ \frac{y_{*1}/\nu_{*1}}{y_{*2}/\nu_{*2}} \le t \frac{\lambda_{*2}}{\lambda_{*1}} \frac{ab}{\nu_{*1}} \frac{\nu_{*2}}{b\nu_e} \right\} \\
&= F_F\left(t \frac{\lambda_{*2}}{\lambda_{*1}} \frac{ab}{\nu_{*1}} \frac{\nu_{*2}}{b\nu_e}; \nu_{*1}, \nu_{*2}, \omega_* \right) \\
&= F_F(t; \nu_{*1}, \nu_{*2}, \omega_*) \, .
\end{aligned}$$

□

Lemma 4.3 For the nonnull case, a ratio estimating ϵ_n in terms of correlated, but unbiased, estimators is

$$\widetilde{\epsilon}_n = \frac{\nu_e(\nu_e+1)\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2\nu_e \text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) + 2[\nu_e(\nu_e+1)-2]\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\left\{ \nu_e^2 \text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e \text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) + 2[\nu_e(\nu_e+1)-2]\text{tr}(\widehat{\boldsymbol{\Sigma}}_* \boldsymbol{\Delta}/a) \right\}} \, . \tag{B.4}$$

$$\widetilde{\epsilon}_r = \frac{(\nu_e+1)b\widehat{\epsilon} - 2}{b(\nu_e - b\widehat{\epsilon})} = \widetilde{\epsilon}_n | \boldsymbol{\Delta} = \mathbf{0} \, . \tag{B.5}$$

This approximately unbiased estimator reduces to the rank-adjusted sphericity estimator under the null case as depicted above in equation B.5.

Proof. Unbiased estimators for $\text{tr}^2(\boldsymbol{\Sigma}_*)$ and $\text{tr}(\boldsymbol{\Sigma}_*^2)$ are

$$\widehat{\tau}_1 = [\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2(\nu_e+1)^{-1}\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)]\{1 - 2[\nu_e(\nu_e+1)]^{-1}\}^{-1}$$

$$\widehat{\tau}_2 = \left(\nu_e^2 \text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e \text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) \right)[\nu_e(\nu_e+1)-2]^{-1} \, ,$$

as introduced in Lemma 2.1. As shown below, $\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)$ and $\text{tr}(\widehat{\boldsymbol{\Sigma}}_* \boldsymbol{\Delta})$ are unbiased estimators for $\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)$ and $\text{tr}(\widehat{\boldsymbol{\Sigma}}_* \boldsymbol{\Delta})$, respectively. Thus,

$$
\begin{aligned}
\widetilde{\epsilon}_n &= \frac{\widehat{\tau}_1 + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\Big[\widehat{\tau}_2 + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\Big]} \\
&= \frac{\Big[\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2(\nu_e+1)^{-1}\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)\Big]\{1 - 2[\nu_e(\nu_e+1)]^{-1}\}^{-1} + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\Big\{\Big[\nu_e^2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)\Big][\nu_e(\nu_e+1)-2]^{-1} + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\Big\}} \\
&= \frac{\Big[\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - \frac{2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)}{(\nu_e+1)}\Big]\Big\{1 - \frac{2}{[\nu_e(\nu_e+1)]}\Big\}^{-1} + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\Bigg\{\frac{\Big[\nu_e^2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)\Big]}{[\nu_e(\nu_e+1)-2]} + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\Bigg\}} \\
&= \frac{\Big[\frac{(\nu_e+1)\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)}{(\nu_e+1)}\Big]\Big\{\frac{\nu_e(\nu_e+1)-2}{[\nu_e(\nu_e+1)]}\Big\}^{-1} + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\Bigg\{\frac{\Big[\nu_e^2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)\Big]}{[\nu_e(\nu_e+1)-2]} + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\Bigg\}} \\
&= \frac{\Big[\frac{\nu_e(\nu_e+1)\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2\nu_e\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)}{\nu_e(\nu_e+1)-2}\Big] + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\Bigg\{\frac{\Big[\nu_e^2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)\Big]}{[\nu_e(\nu_e+1)-2]} + 2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\Bigg\}} \\
&= \frac{\nu_e(\nu_e+1)\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2\nu_e\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) + 2[\nu_e(\nu_e+1)-2]\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)\text{tr}(\boldsymbol{\Delta}/a)}{b\Big\{\nu_e^2\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) + 2[\nu_e(\nu_e+1)-2]\text{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\Big\}} \,.
\end{aligned}
$$

In the null case, $\boldsymbol{\Delta} = \mathbf{0}$, and $\widetilde{\epsilon}_n$ reduces to the rank-adjusted sphericity estimator,

$$
\widetilde{\epsilon}_r = \frac{(\nu_e+1)b\widehat{\epsilon} - 2}{b(\nu_e - b\widehat{\epsilon})} = \widetilde{\epsilon}_n|\boldsymbol{\Delta} = \mathbf{0} \,.
$$

□

First Moment Derivations for $\text{tr}(\widehat{\boldsymbol{\Sigma}}_*)$, $\text{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)$, $\text{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)$ and $\text{tr}(\boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_*)$

Following the method presented by Wishart (1928), let $\boldsymbol{S} = \nu\widehat{\boldsymbol{\Sigma}}_* \sim \mathcal{W}_b(\nu, \boldsymbol{\Sigma}_*)$, such that $\nu = N - r$. In general, the notation introduced by Wishart (1928) is followed with the exception of defining $\langle\boldsymbol{\Sigma}\rangle_{jj}$ as σ_{jj}, while Wishart (1928) defines $\langle\boldsymbol{\Sigma}\rangle_{jj}$ as σ_j^2. Note $\rho_{jk} = \sigma_{jk}/(\sigma_{jj}\sigma_{kk})^{1/2}$. A key sentence in Wishart (1928) reads, with emphasis not in the original, "…moment coefficients are in all cases *except the first* calculated about the mean of the sample…". Using the notation $\mu(n)$ to indicate the expression given in equation n at

the end of Wishart's seminal paper,

$$\mathrm{E}[\mathrm{tr}(\boldsymbol{S})] = \mathrm{E}\left[\sum_{j=1}^{b} s_{jj}\right] = \sum_{j=1}^{b} \mathrm{E}[s_{jj}] = \sum_{j=1}^{b} \nu\sigma_{jj} \tag{B.6}$$

such that

$$\mathrm{E}[s_{jj}] = \mu(1)_j = \nu\sigma_{jj} \; . \tag{B.7}$$

Thus,

$$\mathrm{E}\Big[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)\Big] = (1/\nu)\mathrm{E}[\mathrm{tr}(\boldsymbol{S})] = (1/\nu)\sum_{j=1}^{b} \nu\sigma_{jj} = \sum_{j=1}^{b} \sigma_{jj} = \mathrm{tr}(\boldsymbol{\Sigma}_*) \; . \tag{B.8}$$

□

With $\boldsymbol{S}^2 = (\nu\widehat{\boldsymbol{\Sigma}}_*)^2$,

$$\mathrm{E}\big[\mathrm{tr}(\boldsymbol{S}^2)\big] = \mathrm{E}\left[\left(\sum_{j=1}^{b}\sum_{k=1}^{b} s_{jk}^2\right)\right] = \sum_{j=1}^{b}\sum_{k=1}^{b} \mathrm{E}\big(s_{jk}^2\big) \tag{B.9}$$

such that

$$\begin{aligned} \mathrm{E}\big(s_{jk}^2\big) &= \mathrm{E}\big\{[[s_{jk} - \mathrm{E}(s_{jk})] + \mathrm{E}(s_{jk})]^2\big\} \\ &= \mathrm{E}\big\{[s_{jk} - \mathrm{E}(s_{jk})]^2 + 2\mathrm{E}(s_{jk})[s_{jk} - \mathrm{E}(s_{jk})] + [\mathrm{E}(s_{jk})]^2\big\} \\ &= \mathrm{E}\big\{[s_{jk} - \mathrm{E}(s_{jk})]^2\big\} + 2\mathrm{E}(s_{jk})\mathrm{E}[s_{jk} - \mathrm{E}(s_{jk})] + [\mathrm{E}(s_{jk})]^2 \\ &= \mathrm{E}\big\{[s_{jk} - \mathrm{E}(s_{jk})]^2\big\} + [\mathrm{E}(s_{jk})]^2 \; . \end{aligned} \tag{B.10}$$

There are two cases to consider as shown below:

$$\mathrm{E}\big(s_{jk}^2\big) = \begin{cases} \mathrm{E}\big(s_{jj}^2\big) \;=\; \big[s_{jj}^2 - [\mathrm{E}(s_{jj})]^2\big] + [\mathrm{E}(s_{jj})]^2 \;=\; \mu(3)_j + \big[\mu(1)_j\big]^2 & \text{(B.11)} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } j = k & \\ \mathrm{E}\big(s_{jk}^2\big) = \big[s_{jk}^2 - [\mathrm{E}(s_{jk})]^2\big] + [\mathrm{E}(s_{jk})]^2 = \mu(5)_{jk} + \big[\mu(2)_{jk}\big]^2 & \text{(B.12)} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } j \neq k & \end{cases}$$

$$\begin{aligned} \mathrm{E}\big(s_{jj}^2\big) = \mu(3)_j + \big[\mu(1)_j\big]^2 &= 2\nu\sigma_{jj}^2 + (\nu\sigma_{jj})^2 \\ &= \nu\sigma_{jj}^2(2+\nu) \end{aligned} \tag{B.13}$$

$$\begin{aligned} \mathrm{E}\big(s_{jk}^2\big) = \mu(5)_{jk} + \big[\mu(2)_{jk}\big]^2 &= \nu\sigma_{jj}\sigma_{kk}\big(1+\rho_{jk}^2\big) \\ &= \nu\sigma_{jj}\sigma_{kk}\Big(1+\Big[\tfrac{\sigma_{jk}^2}{(\sigma_{jj}\sigma_{kk})}\Big]\Big) \\ &= \nu\sigma_{jj}\sigma_{kk} + \Big[\tfrac{\nu\sigma_{jj}\sigma_{kk}\sigma_{jk}^2}{(\sigma_{jj}\sigma_{kk})}\Big] \\ &= \nu\sigma_{jj}\sigma_{kk} + \nu\sigma_{jk}^2 \\ &= \nu\big(\sigma_{jj}\sigma_{kk} + \sigma_{jk}^2\big) \end{aligned} \tag{B.14}$$

$$\mathrm{E}\big(s_{jk}^2\big) = \begin{cases} \nu\sigma_{jj}^2(2+\nu) & \text{if } j = k \qquad \text{(B.15)} \\ \nu\big(\sigma_{jj}\sigma_{kk} + \sigma_{jk}^2\big) & \text{if } j \neq k\,. \qquad \text{(B.16)} \end{cases}$$

Thus,

$$\mathrm{E}\Big[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)\Big] = \big(1/\nu^2\big)\mathrm{E}\big[\mathrm{tr}(\boldsymbol{S}^2)\big] = \begin{cases} \frac{1}{\nu^2}\sum_{j=1}^{b}\nu\sigma_{jj}^2(2+\nu) & \text{if } j = k \qquad \text{(B.17)} \\ \frac{1}{\nu^2}\sum_{j=1}^{b}\sum_{k=1}^{b}\nu\big(\sigma_{jj}\sigma_{kk} + \sigma_{jk}^2\big) & \text{if } j \neq k\,. \qquad \text{(B.18)} \end{cases}$$

When simplified,

$$\mathrm{E}\Big[\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)\Big] = \big(1/\nu^2\big)\big[\nu(\nu+1)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu\mathrm{tr}^2(\boldsymbol{\Sigma}_*)\big]\,. \tag{B.19}$$

□

With $\boldsymbol{S} = \nu\widehat{\boldsymbol{\Sigma}}_*$,

$$\begin{aligned} \mathrm{E}\big[\mathrm{tr}^2(\boldsymbol{S})\big] &= \mathrm{E}[\mathrm{tr}(\boldsymbol{S})\mathrm{tr}(\boldsymbol{S})] = \mathrm{E}\left[\left(\sum_{j=1}^{b} s_{jj}\right)\left(\sum_{k=1}^{b} s_{kk}\right)\right] \\ &= \mathrm{E}\left[\sum_{j=1}^{b}\sum_{k=1}^{b} s_{jj}s_{kk}\right] = \sum_{j=1}^{b}\sum_{k=1}^{b}\mathrm{E}[s_{jj}s_{kk}] \end{aligned} \tag{B.20}$$

such that

$$\mathrm{E}[s_{jj}s_{kk}] = \mu(4)_{jk} = 2\nu\sigma_{jk}^2 + \nu^2\sigma_{jj}\sigma_{kk}\ . \tag{B.21}$$

Thus,

$$\begin{aligned} \mathrm{E}\Big[\mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)\Big] &= \big(1/\nu^2\big)\mathrm{E}\big[\mathrm{tr}^2(\boldsymbol{S})\big] = \big(1/\nu^2\big)\sum_{j=1}^{b}\sum_{k=1}^{b}\big(2\nu\sigma_{jk}^2 + \nu^2\sigma_{jj}\sigma_{kk}\big) \\ &= \big(1/\nu^2\big)\Big[2\nu\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) + \nu^2\mathrm{tr}^2(\boldsymbol{\Sigma}_*)\Big]\ . \end{aligned} \tag{B.22}$$

□

Finally, for $\boldsymbol{S} = \nu\widehat{\boldsymbol{\Sigma}}_* \sim \mathcal{W}_b(\nu, \boldsymbol{\Sigma}_*)$ and $\boldsymbol{\Delta S} = \nu\boldsymbol{\Phi}'_{\boldsymbol{\Delta}}\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Phi}_{\boldsymbol{\Delta}} \sim \mathcal{W}_{s_*}(\nu, \boldsymbol{\Phi}'_{\boldsymbol{\Delta}}\boldsymbol{\Sigma}_*\boldsymbol{\Phi}_{\boldsymbol{\Delta}})$,

$$\mathrm{E}[\mathrm{tr}(\boldsymbol{\Delta S})] = \mathrm{E}[\mathrm{tr}(\boldsymbol{\Phi}'_{\boldsymbol{\Delta}}\boldsymbol{S}\boldsymbol{\Phi}_{\boldsymbol{\Delta}})] = \mathrm{E}\left[\sum_{j=1}^{s_*} s_{\Delta\Sigma_{jj}}\right] = \sum_{j=1}^{s_*}\mathrm{E}\big[s_{\Delta\Sigma_{jj}}\big] = \sum_{j=1}^{s_*}\nu\sigma_{jj} \tag{B.23}$$

such that

$$\mathrm{E}\big[s_{\Delta\Sigma_{jj}}\big] = \mu(1)_j = \nu\sigma_{jj}\ . \tag{B.24}$$

Thus,

$$\begin{aligned} \mathrm{E}\Big[\mathrm{tr}(\boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_*)\Big] &= \mathrm{E}[\mathrm{tr}(\boldsymbol{\Delta S})]/\nu = (1/\nu)\sum_{j=1}^{s_*}\nu\sigma_{jj} = \sum_{j=1}^{s_*}\sigma_{jj} = \mathrm{tr}(\boldsymbol{\Phi}'_{\boldsymbol{\Delta}}\boldsymbol{\Sigma}_*\boldsymbol{\Phi}_{\boldsymbol{\Delta}}) \\ &= \mathrm{tr}(\boldsymbol{\Delta\Sigma}_*)\ . \end{aligned} \tag{B.25}$$

□

Thus, $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)$ and $\mathrm{tr}(\boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_*)$ are unbiased estimators, while $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)$ and $\mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)$ are biased estimators.

Appendix C: Simulation Details

Covariance Conditions

Covariance conditions 5-8 from Table III of Coffey and Muller (2003) were used for each example described below: $\boldsymbol{\Sigma}_* = \text{Dg}(\boldsymbol{\lambda}_j)$ for $j \in \{1, 2, 3, 4\}$, with

$$\begin{aligned}\boldsymbol{\lambda}_1' &= [0.47960\ 0.01000\ 0.01000\ 0.01000]\,, \boldsymbol{\lambda}_2' = [0.34555\ 0.06123\ 0.05561\ 0.04721]\,, \\ \boldsymbol{\lambda}_3' &= [0.23555\ 0.17123\ 0.05561\ 0.04721]\,, \boldsymbol{\lambda}_4' = [0.12740\ 0.12740\ 0.12740\ 0.12740]\,.\end{aligned} \tag{C.1}$$

Thus, $\epsilon \in \{0.28, 0.51, 0.72, 1.00\}$. Given $\boldsymbol{\Sigma}_* = \text{Dg}(\boldsymbol{\lambda}_j)$, it follows that $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Sigma}_*\boldsymbol{U}'$.

Test of Interaction with rank$(\boldsymbol{X}) > 1$ Example

The cases consisted of 5 repeated measures, $N \in \{16, 32, 48\}$, and rank$(\boldsymbol{X}) \in \{2, 4, 8, 16\}$. For obvious reasons, a rank of $\boldsymbol{X}$ equal to 16 was not considered for the smallest sample size. All four covariance patterns were factorially combined with the sample sizes and ranks $\boldsymbol{X}$. In the multivariate model,

$$\underset{(N \times 5)}{\boldsymbol{Y}} = \underset{(N \times q \times 5)}{\boldsymbol{X}\boldsymbol{B}} + \underset{(N \times 5)}{\boldsymbol{E}}\,, \tag{C.2}$$

$\boldsymbol{X} = \boldsymbol{I}_q \otimes \mathbf{1}_{repn}$, such that $repn = N/q$, and $\otimes$ is a Kronecker product. If

$$\boldsymbol{I}_{16} = \begin{bmatrix} \underset{q \times 5}{\boldsymbol{I}_a} & \underset{q \times 11}{\boldsymbol{I}_b} \\ \underset{(16-q) \times 5}{\boldsymbol{I}_c} & \underset{(16-q) \times 11}{\boldsymbol{I}_d} \end{bmatrix}, \tag{C.3}$$

$\boldsymbol{B} = \beta_P \cdot \boldsymbol{I}_a$, such that β_P was the scaling factor for $\boldsymbol{B}$ corresponding to approximate target power $P \in \{0.20, .0.50, 0.80\}$, using methods for the rank-adjusted Huynh-Feldt power approximation as presented in section 3.3. The within-subject contrast, $\boldsymbol{U}$, (5×4), was an orthonormal trends matrix for linear, quadratic, cubic and quartic trends,

$$\boldsymbol{U} = \begin{bmatrix} -2/\sqrt{10} & 2/\sqrt{14} & -1/\sqrt{10} & 1/\sqrt{70} \\ -1/\sqrt{10} & -1/\sqrt{14} & 2/\sqrt{10} & -4/\sqrt{70} \\ 0/\sqrt{10} & -2/\sqrt{14} & 0/\sqrt{10} & 6/\sqrt{70} \\ 1/\sqrt{10} & -1/\sqrt{14} & -2/\sqrt{10} & -4/\sqrt{70} \\ 2/\sqrt{10} & 2/\sqrt{14} & 1/\sqrt{10} & 1/\sqrt{70} \end{bmatrix}. \tag{C.4}$$

The between-subject contrast, $\boldsymbol{C}$, is a $(q-1 \times q)$ orthonormal trends matrix up to the $(q-1)$ order across rows, similar to the within-subject contrast matrix is across columns. The contrasts yield a test of interaction of between- and within-subject trends. Without loss of generality, assume $\boldsymbol{\Theta}_0 = \mathbf{0}$. A test size, α, of 0.05 was used.

CLAHE Mammography Example

To improve contrast in digital mammography, computer scientists developed the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm. Independent observers considered $3 \times 3 = 9$ Clip $\times$ Region combinations for one group only. Region denotes the size of the image (pixels^2) at which contrasts are controlled and Clip level limits the maximum contrast adjustment. In the multivariate model $\boldsymbol{X} = \mathbf{1}_N$, while within-person factors Clip and Region gave $\boldsymbol{Y}$, $(N \times 9)$. Also $\boldsymbol{B}$, (1×9), contained mean $\log_{10}(\text{contrast})$ for the unprocessed condition minus the mean for each of the nine combinations of Clip and Region ($\beta_{\text{cr}} = \mu_{\text{unprocessed}} - \mu_{\text{cr}}$). If $\boldsymbol{T}_{\text{c}}$ contains orthonormal linear and quadratic trends for $\log_2(\text{Clip}) \in \{1, 2, 4\}$, and $\boldsymbol{T}_{\text{r}}$ does the same for $\log_2(\text{Region}) \in \{1, 3, 5\}$, then the 9×4 within-persons contrast matrix, $\boldsymbol{U}_{\text{cr}}$ is

$$\boldsymbol{U}_{\text{cr}} = \boldsymbol{T}_{\text{c}} \otimes \boldsymbol{T}_{\text{r}} = \begin{bmatrix} -4/\sqrt{42} & 2/\sqrt{14} \\ -1/\sqrt{42} & -3/\sqrt{14} \\ 5/\sqrt{42} & 1/\sqrt{14} \end{bmatrix} \otimes \begin{bmatrix} -1/\sqrt{2} & 1/\sqrt{6} \\ 0 & -2/\sqrt{6} \\ 1/\sqrt{2} & 1/\sqrt{6} \end{bmatrix}. \qquad \text{(C.5)}$$

With L the linear and Q the quadratic trends for interaction components being tested, $\boldsymbol{U}_{\text{cr}} = [\, \boldsymbol{u}_{\text{LL}} \quad \boldsymbol{u}_{\text{LQ}} \quad \boldsymbol{u}_{\text{QL}} \quad \boldsymbol{u}_{\text{QQ}} \,]$.

All four covariance patterns were factorially combined with $N \in \{10, 20, 40\}$. The multivariate test considered $\boldsymbol{\Theta}_{\text{cr}} = \beta_P \cdot [\, 0.5 \; 1.0 \; -1.0 \; 0.5 \,]$ with a test size, α, of 0.05, with β_P the scaling factor for $\boldsymbol{B}$ corresponding to approximate target power $P \in \{0.20, .0.50, 0.80\}$ for the Geisser-Greenhouse approximation using methods presented in Muller *et al.* (2007). Simulated population power values were computed for 500,000 replications per condition. The conditions set forth in this example were utilized in

section 4.6.1. Further description of the CLAHE Mammography example may be found in Muller *et al.* (2007).

Standard Error Calculations

Standard error values are presented throughout the document in each table in which observed mean simulation values are tabulated. When calculating the standard error of observed mean power values, the equation providing the most conservative value was used, $S.E. = \sqrt{(0.50)(0.50)/N_{rep}}$, such that N_{rep} was the number of replications used for each case in a simulation. For similar calculations for standard errors of the observed mean sphericity estimates, the interaction test sizes and the confidence interval coverages, a more liberal, but, simultaneously, more appropriate equation was used, $S.E. = \sqrt{(0.95)(0.05)/N_{rep}}$. In Chapter 4, 95% Score half confidence intervals equal to $1.96 \times S.E.$ are provided for the simulated confidence interval coverage values.

Computational Methods

All power computations were conducted in SAS/IML (SAS 9.1, SAS Institute, Copyright 2003). Free software LINMOD 3.4 (http://ehpr.ufl.edu/muller/) was used for all data analysis and includes new methods. Free software POWERLIB 2.1 (http://ehpr.ufl.edu/muller/) was used for all power analysis and includes the new methods.

References

Barton, C. N. and Cramer, E. C. (1989). Hypothesis testing in multivariate linear models with randomly missing data. *Communications in Statistics*, **18**, 875-895.

Beale, E. M. L. and Little, R. J. A. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, **37**, 129-145.

Boik, R. J. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. *Psychometrika*, **46**, 241-255.

Box, G. E. P. (1954a). Some theorems on quadratic forms applied in the study of analysis of variance problems: I. Effects of inequality of variance in a one-way classification. *Annals of Mathematical Statistics*, **25**, 290-302.

Box, G. E. P. (1954b). 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. *Annals of Mathematical Statistics*, **25**, 484-498.

Catellier, D. J. and Muller, K. E. (2000). Tests for Gaussian repeated measures with missing data in small samples. *Statistics in Medicine*, **19**, 1101-1114.

Coffey, C. S. and Muller, K. E. (2003). Properties of internal pilots with the univariate approach to repeated measures. *Statistics in Medicine*, 22, 2469-2485.

Cole, J. W. L. and Grizzle, J. E. (1966). Applications of multivariate analysis of variance to repeated measurements experiments. *Biometrics*, **22**, 810-828.

Constantine, A. G. (1963). Some noncentral distribution problems in multivariate analysis. *Annals of Mathematical Statistics*, **34**, 1270-1285.

Davidson, M. L. (1972). Univariate versus multivariate tests in repeated measures experiments. *Psychological Bulletin*, **77**, 446-452.

Davies, R. B. (1980). Distribution of a linear combination of chi-square random variables. *Applied Statistics*, **29**, 323-333.

Dudewicz, E. J. (1972). Confidence intervals for power with special reference to medical trials. *Australian Journal of Statistics*, **14**, 211-216.

Geisser, S. and Greenhouse, S. W. (1958). An extension of Box's results on the use of the F distribution in multivariate analysis. *Annals of Mathematical Statistics*, **29**, 885-891.

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

Gueorguieva, R. and Krystal, J. H. (2004). Move over ANOVA. *Arch General Psychiatry*, **61**, 310-317.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika*, **54**, 93-108.

Huynh, H. (1978). Some approximate tests for repeated measurement designs. *Psychometrika*, **43**, 161-175.

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

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

Huynh, H. and Feldt, L. S. (1980). Performance of traditional F tests in repeated measures designs under variance heterogeneity. *Communications in Statistics: Series A*, **9**, 61-74.

Huynh, H. and Mandeville, G. (1979). Validity conditions in repeated measures designs. *Psychological Bulletin*, 86, 964-973.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrika*, **42**, 805-820.

John, S. (1971). Some optimal multivariates test. *Biometrika*, **58**, 123-127.

Kim, H. Y., Gribbin, M. J., Muller, K. E. and Taylor D. J. (2006). Analytic, computational and approximate forms for ratios of noncentral and central Gaussian quadratic forms. *Journal of Computational and Graphical Statistics*, **15**, 443-459.

Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin*, **92**, 513-516.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963-974.

Lee, Y. S. (1971). Asymptotic formulae for the distribution of the multivariate test statistic: Power comparisons of certain multivariate tests. *Biometrika*, **58**, 647-651.

Morrison, D. F. (1990). *Multivariate Statistical Methods*, 3rd ed., New York: McGraw-Hill.

Muller, K. E. and Barton, C. N. (1989). Approximate power for repeated measures ANOVA lacking sphericity. *Journal of the American Statistical Association*, 84, 549-555. Also see (1991). Correction to Approximate power for repeated-measures ANOVA lacking sphericity. *Journal of the American Statistical Association*, **86**, 255-256.

Muller, K. E. and Benignus, V. A. (1992). Increasing scientific power with statistical power. *Neurotoxicology and Teratology*, **14**, 211-219.

Muller, K. E., Edwards, L. J., Simpson, S. L., and Taylor, D. J. (2007). Statistical tests with accurate size and power for balance linear mixed models, Statistics in Medicine, *in press*.

Muller K. E. and Fetterman B. A. (2002). *Regression and ANOVA: An Integrated Approach Using SAS® Software*. Cary, NC: SAS Institute, Chapter 17.

Muller, K. E., LaVange, L. M., Ramey, S. L., and Ramey, C. T. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association*, **87**, 1209-1226.

Muller, K. E. and Peterson, B. L. (1984). Practical methods for computing power in testing the multivariate general linear hypothesis. *Computational Statistics & Data Analysis*, **2**, 143-158.

Muller, K. E. and Stewart, P. W. (2006). *Linear Model Theory; Univariate, Multivariate and Mixed Models*, New York: Wiley.

Nagarsenker, B. N. and Suniaga, J. (1983). Distributions of a class of statistics useful in multivariate analysis. *Journal of the American Statistical Association*, **78**, 472-475.

O'Brien, R. G. and Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, **97**, 316-333.

Olson, C. L. (1974). Comparative robustness of six tests in multivariate analysis of variance. *Journal of the American Statistical Association*, **69**, 894-908.

Olson, C. L. (1976). Choosing a test statistic in multivariate analysis. *Psychological Bulletin*, **83**, 579-586.

Olson, C. L. (1979). Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens. *Psychological Bulletin*, **86**, 1350-1352.

Posten, H. O. and Bargmann, R. E. (1964). Power of the likelihood ratio test of the general linear hypothesis in multivariate analysis. *Biometrika*, **51**, 467-480.

Rao, C. R. (1972). Recent trends of research work in multivariate analysis. *Biometrics*, **28**, 3-22.

Rouanet, H. and Lepine, D. (1970). Comparison between treatments in a repeated measures design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, **23**, 147-163.

Schatzoff, M. (1966). Sensitivity comparisons among tests of the general linear hypothesis. *Journal of the American Statistical Association*, **61**, 415-435.

Smith, H., Gnanadesikan, R. and Hughes, J. B. (1962). Multivariate Analysis of Variance (MANOVA). *Biometrics*, **18**, 22-41.

Stevens, J. P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin*, **88**, 728-737.

Sugiura, N. and Fujikoshi, Y. (1969). Asymptotic expansion of the non-null distributions on the likelihood ratio criteria for multivariate linear hypothesis and independence. *Annals of Mathematical Statistics*, **40**, 942-952.

Taylor, D. J. and Muller, K. E. (1995). Computing confidence bounds for power and sample size of the general linear univariate model. *American Statistician*, **49**, 43-47.

Taylor, D. J. and Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics: Theory and Methods*, **25**, 1595-1610.

Venables, W. (1975). Calculation of confidence intervals for noncentrality parameters. *Journal of the Royal Statistical Society. Series B (Methodological)*, **37**, 406-412.

Wallenstein, S. and Fleiss, J. L. (1979). Repeated measures analysis of variance when the correlations have a certain pattern. *Psychometrika*, **44**, 229-233.

Wishart, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. *Biometrika*, **20A**, 32-52.