BMC
Medical Research Methodology

RESEARCH ARTICLE | Open Access

# Confidence regions for repeated measures ANOVA power curves based on estimated covariance

Matthew J Gribbin[1*], Yueh-Yun Chi[2], Paul W Stewart[3] and Keith E Muller[4]

**Abstract**

**Background:** Using covariance or mean estimates from previous data introduces randomness into each power value in a power curve. Creating confidence intervals about the power estimates improves study planning by allowing scientists to account for the uncertainty in the power estimates. Driving examples arise in many imaging applications.

**Methods:** We use both analytical and Monte Carlo simulation methods. Our analytical derivations apply to power for tests with the univariate approach to repeated measures (UNIREP). Approximate confidence intervals and regions for power based on an estimated covariance matrix and fixed means are described. Extensive simulations are used to examine the properties of the approximations.

**Results:** Closed-form expressions are given for approximate power and confidence intervals and regions. Monte Carlo simulations support the accuracy of the approximations for practical ranges of sample size, rank of the design matrix, error degrees of freedom, and the amount of deviation from sphericity. The new methods provide accurate coverage probabilities for all four UNIREP tests, even for small sample sizes. Accuracy is higher for higher power values than for lower power values, making the methods especially useful in practical research conditions. The new techniques allow the plotting of power confidence regions around an estimated power curve, an approach that has been well received by researchers. Free software makes the new methods readily available.

**Conclusions:** The new techniques allow a convenient way to account for the uncertainty of using an estimated covariance matrix in choosing a sample size for a repeated measures ANOVA design. Medical imaging and many other types of healthcare research often use repeated measures ANOVA.

**Keywords:** Sample size, Replication study, Study planning, Univariate approach, UNIREP

## Background

### Motivation

Computing power for a linear model involving repeated measures requires specifying a set of means and a covariance matrix. Scientists usually feel comfortable specifying a pattern of means that corresponds to a difference of clinical or scientific importance. However, specifying plausible variance and covariance values usually requires estimates from a previous study.

Using data from a previous study to estimate the covariance matrix makes the power value a random variable. Kraemer, et al. [1] noted that if the estimated variance is too small, i.e., when the pilot study is overly favorable, power will be over-estimated. Conversely, if the estimated variance is too large (pilot study overly conservative), power will be under-estimated. Maxwell [2] conducted simulations to illustrate the amount of bias that can occur. Taylor and Muller [3] and Muller and Pasour [4] derived exact distributions of noncentrality and power in univariate linear models based on all combinations of estimated variance and means. The results account for power computed conditional on a previous result being significant, or conditional on a previous result being non-significant. The former creates optimistic bias, while the latter creates pessimistic bias.

Providing confidence intervals to account for the uncertainty inherent in the random power values would be

*Correspondence: mgribbin@gmail.com
[1] Department of Biostatistics, MedImmune, Gaithersburg, MD, USA
Full list of author information is available at the end of the article

useful for study planning. For example, a lower bound for power would allow stating that a test has power of at least "*P*" to detect an effect, with a specified confidence. A confidence region for a power curve would be even more informative.

Medical imaging research motivated the work here because it often generates the type of complete data that can be handled with the univariate approach to repeated measures (UNIREP). Muller, et al. [5] reviewed the advantages gained by being able to use the UNIREP model, a special case of the general linear mixed model. The same authors described accurate and convenient power approximations for UNIREP analysis. The four UNIREP tests, Box conservative, Geisser-Greenhouse, Huynh-Feldt, and the uncorrected, all use the same test statistic. For data analysis, UNIREP tests differ only by their respective degrees of freedom due to different degrees of freedom multipliers, which measure sphericity in the error covariance for the hypothesis variables. Muller and Stewart [6] provided detailed discussion of the basic theory for both the null and non-null cases. Earlier work detailed basic UNIREP theory. Box [7,8], Geisser and Greenhouse [9,10] and Huynh and Feldt [11] gave null results. Davies [12] and Muller and Barton [13,14] treated the non-null case.

Browne [15] evaluated the impact of using a pilot study to estimate the variance for a *t*-test. More generally, Taylor and Muller [16] demonstrated how to construct exact power confidence intervals for the general linear univariate model for a data-estimated variance and fixed means. The same authors also generalized the result to provide an exact confidence region around a power curve. Parallel results for the UNIREP setting would be equally useful. We generalize the methods in Taylor and Muller [16] to UNIREP tests for repeated measures. We use analytic and simulation results to demonstrate that the techniques allow computing approximate confidence intervals and regions for power with good accuracy for the UNIREP tests, based on an estimated covariance matrix and fixed means.

## Existing results

A vector $\boldsymbol{z}$, $(n \times 1)$, is lower case bold. A matrix, $\boldsymbol{Z}$, is upper case bold with transpose $\boldsymbol{Z}'$, inverse $\boldsymbol{Z}^{-1}$ and generalized inverse $\boldsymbol{Z}^{-}$. Also, $\boldsymbol{1}_n$ 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 $z_i$ is written $\mathrm{Dg}(\boldsymbol{z})$. The expected value, variance, and trace are $\mathrm{E}(\boldsymbol{Z})$, $\mathcal{V}(\boldsymbol{Z})$, and $\mathrm{tr}(\boldsymbol{Z})$, respectively. Throughout, $Z \sim \chi^2(\nu, \omega)$ indicates that $Z$ has a noncentral chi-square distribution with $\nu$ degrees of freedom and noncentrality $\omega$, while $Z \sim \chi^2(\nu)$ indicates a central distribution. Similarly, $Z \sim F(\nu_1, \nu_2, \omega)$ indicates $X$ has a noncentral $F$ distribution with $\nu_1$ numerator and $\nu_2$ denominator degrees of freedom, and noncentrality $\omega$ with cumulative distribution function $F_F(\nu_1, \nu_2, \omega)$. A central $F$ is written $Z \sim F(\nu_1, \nu_2)$ with quantile $q$ indicated $F_F^{-1}(q; \nu_1, \nu_2)$. Writing $\boldsymbol{z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ indicates $\boldsymbol{z}$ $(p \times 1)$ is Gaussian with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ $(p \times p)$. If $\boldsymbol{Z}$ $(N \times p)$ has independent rows and $[\mathrm{row}_i(\boldsymbol{Z})]' \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, then $\boldsymbol{S} = \boldsymbol{Z}'\boldsymbol{Z} \sim \mathcal{W}_p(N, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$ indicates $\boldsymbol{S}$ follows a Wishart distribution with $N$ degrees of freedom, covariance $\boldsymbol{\Sigma}$, and noncentrality $\boldsymbol{\Omega} = \mathrm{E}(\boldsymbol{Z}')\,\mathrm{E}(\boldsymbol{Z})\,\boldsymbol{\Sigma}^{-1}$.

The general linear multivariate model,

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

assumes $N$ independent rows and $[\mathrm{row}_i(\boldsymbol{Y})]' \sim \mathcal{N}_p\left([\mathrm{row}_i(\boldsymbol{X})\,\boldsymbol{B}]', \boldsymbol{\Sigma}\right)$. In the model, $\boldsymbol{X}$ is the fixed, known design matrix with $1 \leq \mathrm{rank}(X) \leq q$, and $\boldsymbol{B}$ contains fixed, unknown regression coefficients. For repeated measures ANOVA, one-group designs have $\mathrm{rank}(\boldsymbol{X}) = 1$, and two-group comparisons have $\mathrm{rank}(\boldsymbol{X}) = 2$. The associated general linear hypothesis is

$$H_0: \underset{(a \times b)}{\boldsymbol{\Theta}} = \underset{(a \times q)\,(q \times p)\,(p \times b)}{\boldsymbol{CBU}} = \underset{(a \times b)}{\boldsymbol{\Theta}_0}, \qquad (2)$$

such that $\boldsymbol{C}$ defines the between-subject effects (rank $a$) while $\boldsymbol{U}$ defines the within-subject effects (rank $b$). Requiring estimable $\boldsymbol{\Theta}$ and full rank $\{\boldsymbol{C}, \boldsymbol{U}\}$ ensure a testable hypothesis. Appropriate selections of the contrast matrices ($\boldsymbol{C}$ and $\boldsymbol{U}$) and null matrix ($\boldsymbol{\Theta}_0$) allows testing important one-degree-of-freedom parameters, such as the difference between two means, or a comparison of two trends.

For $\boldsymbol{M} = \boldsymbol{C}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{C}'$, unscaled noncentrality is $\boldsymbol{\Delta} = (\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)'\,\boldsymbol{M}^{-1}(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)$, scaled noncentrality is $\boldsymbol{\Omega} = \boldsymbol{\Delta}\boldsymbol{\Sigma}_*^{-1}$. Here $\boldsymbol{\Sigma}_* = \boldsymbol{U}'\boldsymbol{\Sigma}\boldsymbol{U} = \boldsymbol{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})\,\boldsymbol{\Upsilon}'$ is the covariance matrix among the hypothesis variables, with $\boldsymbol{\Upsilon}\boldsymbol{\Upsilon}' = \boldsymbol{\Upsilon}'\boldsymbol{\Upsilon} = \boldsymbol{I}_b$, and $\boldsymbol{\lambda} = \{\lambda_k\}$ the eigenvalues. Estimates are $\widetilde{\boldsymbol{B}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{X}'\boldsymbol{Y}$ and $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{Y}'\left[\boldsymbol{I} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-}\boldsymbol{X}'\right]\boldsymbol{Y}/\nu_e$, with $\nu_e = N - \mathrm{rank}(\boldsymbol{X})$, the error degrees of freedom. Furthermore, $\widehat{\boldsymbol{\Theta}} = \boldsymbol{C}\widetilde{\boldsymbol{B}}\boldsymbol{U}$, $\widehat{\boldsymbol{\Delta}} = (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0)'\boldsymbol{M}^{-1}(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0) \sim \mathcal{W}_b(a, \boldsymbol{\Sigma}_*, \boldsymbol{\Omega})$ 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. The notation follows that in Muller and Stewart [6]. Additional notation is in Appendix A.

The univariate approach to repeated measures can be expressed in terms of the general linear multivariate model. The Box conservative (Box), the Geisser-Greenhouse (GG), the Huynh-Feldt (HF), and the uncorrected (Un) UNIREP tests use the same test statistic,

$$T_u = \frac{\mathrm{tr}(\widehat{\boldsymbol{\Delta}})/a}{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)}, \tag{3}$$

and a central $F$ distribution to approximate the null distribution of $T_u$,

$$\Pr\{T_u \le t\} \approx F_F\left(t; ab\epsilon, \nu_e b\epsilon, 0\right). \tag{4}$$

The sphericity parameter, $\epsilon = \mathrm{tr}^2(\boldsymbol{\Sigma}_*)/[\,b\mathrm{tr}(\boldsymbol{\Sigma}_*^2)]$, quantifies the spread of population eigenvalues and is used to discount the degrees of freedom. The term *sphericity* reflects the fact that uncorrelated Gaussian variables with equal variances in three dimensions have a spherical scattergram. The eigenvalues of $\boldsymbol{\Sigma}_*$ are the variances of the (uncorrelated) principal components of the hypothesis response variables. Perfect sphericity requires $\epsilon = 1$, which occurs with all eigenvalues equal. Minimal sphericity has $\epsilon = 1/b$, which occurs with one nonzero eigenvalue. Other patterns of $\boldsymbol{\Sigma}_*$ have $1/b < \epsilon < 1$.

The Box conservative test uses the fixed, lower bound of $\epsilon$, while the uncorrected test uses the fixed, upper bound of $\epsilon$. With sphericity ($\epsilon = 1$), the uncorrected test is exact and uniformly most powerful (among similarly invariant tests). The Geisser-Greenhouse and Huynh-Feldt tests use the observed data to estimate $\epsilon$. The Geisser-Greenhouse estimator, $\widehat{\epsilon} = \mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)/b\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)$, is the maximum likelihood (ML) estimator. The Huynh-Feldt estimator, $\widetilde{\epsilon} = (Nb\widehat{\epsilon} - 2)\,/\,[b\,(\nu_e - b\widehat{\epsilon})]$ was proposed as the ratio of two unbiased estimators. Their claim holds only for the special case of $\mathrm{rank}(\boldsymbol{X}) = 1$. Lecoutre [17] provided a more general form. In turn, Gribbin [18] and Chi et al. [19] described a rank-adjusted approximately unbiased estimator, $\widetilde{\epsilon}_r = [(\nu_e + 1)\,b\widehat{\epsilon} - 2]\,/\,[b\,(\nu_e - b\widehat{\epsilon})]$, which applies to any general linear multivariate model. 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 (Chi et al. [19]). Only the rank-adjusted Huynh-Feldt estimator will be considered in the remainder of the paper.

Although the four UNIREP tests all use the same test statistic, they each use a different measure of sphericity, here indicated $e$. For data analysis, all four tests use a critical value $q(e) = F_F^{-1}\left(1 - \alpha, \nu_1 e, \nu_2 e\right)$. Here $\nu_1 = ab$ and $\nu_2 = b\nu_e$. The Box test uses $e = 1/b$, the GG test uses $e = \widehat{\epsilon}$, HF uses $e = \tilde{\epsilon}$, and the uncorrected test uses $e = 1$. The p-value is then computed, for observed test statistic $t$, as $p = 1 - F\left(t, \nu_1 e, \nu_2 e\right)$. In all cases $1/b \le \widehat{\epsilon} \le \tilde{\epsilon} \le 1$. In turn, the $p$-values always have the reverse order, with the Box $p$-value being largest, and the uncorrected being smallest.

Muller et al. [5] showed that the distribution function of the UNIREP test statistic can be expressed exactly in terms of the distribution function of the sum of $b$ positively and $b$ negatively weighted independent chi-squares, namely $y_{kh} \sim \chi^2\left(a, \omega_k\right)$ and $y_{ke} \sim \chi^2\left(\nu_e\right)$,

$$\begin{aligned}\Pr\{T_u \le t\} &= \Pr\left\{\frac{\mathrm{tr}(\widehat{\boldsymbol{\Delta}})/a}{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)} \le t\right\} \\ &= \Pr\left\{\sum_{k=1}^{b}\lambda_k y_{kh} - (ta/\nu_e)\sum_{k=1}^{b}\lambda_k y_{ke} \le 0\right\}\end{aligned} \tag{5}$$

Muller et al. [5] also reported accurate $F$ approximations of the form

$$\begin{aligned}\Pr\{T_u \le t\} &\approx \Pr\left\{\frac{\lambda_{*1}y_{*1}/(ab)}{\lambda_{*2}y_{*2}/\left(b\nu_e\right)} \le t\right\} \\ &= F_F\left(t\frac{\lambda_{*2}}{\lambda_{*1}}\frac{ab}{\nu_{*1}}\frac{\nu_{*2}}{b\nu_e}; \nu_{*1}, \nu_{*2}, \omega_*\right).\end{aligned} \tag{6}$$

Here, $y_{*1} \sim \chi^2\left(\nu_{*1}, \omega_*\right)$, $y_{*2} \sim \chi^2\left(\nu_{*2}\right)$, $\mathrm{tr}(\widehat{\boldsymbol{\Delta}}) \approx \lambda_{*1}y_{*1}$ and $\nu_e\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*) \approx \lambda_{*2}y_{*2}$. Parameters $\lambda_{*1}$, $\nu_{*1}$, $\omega_*$, $\lambda_{*2}$, and $\nu_{*2}$ are defined in Appendix A. Power analysis involves $\{\lambda_k\}$ and $\{\omega_k\}$, with

$$\omega_k = \boldsymbol{v}_k'\boldsymbol{\Delta}\boldsymbol{v}_k/\lambda_k, \tag{7}$$

the diagonal elements of the scaled noncentrality, $\boldsymbol{\Omega}_* = \boldsymbol{\Upsilon}'\boldsymbol{\Delta}\boldsymbol{\Upsilon}\mathrm{Dg}(\boldsymbol{\lambda})^{-1} = \boldsymbol{\Delta}_*\mathrm{Dg}(\boldsymbol{\lambda})^{-1}$.

## Methods

### Estimating approximate UNIREP power with estimated covariance and fixed means

By extending results in Muller et al. [5], the following lemma helps simplify the $F$ approximations. Appendix B contains all proofs.

**Lemma 1.** The constant in the critical value of the UNIREP test statistic approximation introduced by Muller et al. [5] is equal to 1,

$$\frac{\lambda_{*2}}{\lambda_{*1}}\frac{ab}{\nu_{*1}}\frac{\nu_{*2}}{b\nu_e} = 1. \tag{8}$$

Thus,

$$\Pr\{T_u \leq t\} \approx \Pr\left\{\frac{\lambda_{*1}y_{*1}/(ab)}{\lambda_{*2}y_{*2}/\left(b\nu_e\right)} \leq t\right\} = F_F\left(t; \nu_{*1}, \nu_{*2}, \omega_*\right) . \tag{9}$$

For known covariance and means, the power approximations for the Box, Geisser-Greenhouse, rank-adjusted Huynh-Feldt, and uncorrected tests are all of the form

$$P = 1-F_F\left[F_F^{-1}(1-\alpha; e_1\cdot ab, e_2\cdot b\nu_e); e_3\cdot ab, e_4\cdot \nu_e b, \frac{\mathrm{tr}\left(\mathbf{\Delta}\right)}{\overline{\lambda}/e_5}\right]. \tag{10}$$

Here, $\overline{\lambda}$ is equal to $\mathrm{tr}\left(\mathbf{\Sigma}_*\right)/b$ with $b$ equal to the rank of $\mathbf{\Sigma}_*$, and $\omega_* = \mathrm{tr}\left(\mathbf{\Delta}\right)/\left(\overline{\lambda}/e_5\right)$. Table 1 contains values for $e_1$ through $e_5$ for the four UNIREP tests when $\epsilon_d = \epsilon = \mathrm{tr}^2(\mathbf{\Sigma}_*)/b\mathrm{tr}(\mathbf{\Sigma}_*^2)$, and $\epsilon_n = \left[\mathrm{tr}^2(\mathbf{\Sigma}_*) + 2\mathrm{tr}(\mathbf{\Sigma}_*)\mathrm{tr}\left(\mathbf{\Delta}/a\right)\right]\left\{b\left[\mathrm{tr}(\mathbf{\Sigma}_*^2) + 2\mathrm{tr}(\mathbf{\Sigma}_*\mathbf{\Delta}/a)\right]\right\}$. The expressions for $\epsilon_d$ and $\epsilon_n$ were derived using the properties described in Lemma B.1.

In practice, some elements of $\{e_1, e_2, e_3, e_4, e_5, \mathrm{tr}(\mathbf{\Delta}), \overline{\lambda}\}$ may be estimated and hence random. The random elements imply random power values, as with estimated covariance and fixed means, $\{\widehat{\mathbf{\Sigma}}_*, \mathbf{\Delta}\}$, for $\widehat{\mathbf{\Sigma}}_* = \widehat{\boldsymbol{E}}'\widehat{\boldsymbol{E}}/\nu_{\mathrm{est}}$ , the unbiased restricted maximum likelihood (REML) 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 $N_{\mathrm{est}}$, design matrix rank of $\mathrm{rank}(\boldsymbol{X}_{\mathrm{est}})$, and $\nu_{\mathrm{est}} = N_{\mathrm{est}} - \mathrm{rank}\left(\boldsymbol{X}_{\mathrm{est}}\right)$ degrees of freedom. The target study for which power is desired has sample size $N$, $\mathrm{rank}(\boldsymbol{X})$ and $\nu_e = N - \mathrm{rank}\left(\boldsymbol{X}\right)$ degrees of freedom.

The ML estimator from the Geisser-Greenhouse test, $\widehat{\epsilon} = \mathrm{tr}^2(\widehat{\mathbf{\Sigma}}_*)/[b\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*^2)]$, is an obvious estimator for the target study's $\epsilon$. For power analysis, a parallel estimator is available for $\epsilon_n$:

$$\widehat{\epsilon}_n = \frac{\mathrm{tr}^2(\widehat{\mathbf{\Sigma}}_*) + 2\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*)\mathrm{tr}\left(\mathbf{\Delta}/a\right)}{b\left[\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*^2) + 2\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*\mathbf{\Delta}/a)\right]}. \tag{11}$$

A better choice, given in the following lemma, uses a ratio of unbiased estimators. The result generalizes the rank-adjusted Huynh-Feldt estimator for data analysis. Appendix B has derivations of moments as well as all proofs.

**Lemma 2.** For the non-null case, a ratio estimating $\epsilon_n$ in terms of correlated, but unbiased, estimators is

$$\widetilde{\epsilon}_n = \frac{\nu_{\mathrm{est}}\left(\nu_{\mathrm{est}}+1\right)\mathrm{tr}^2(\widehat{\mathbf{\Sigma}}_*) - 2\nu_{\mathrm{est}}\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*^2) + 2\left[\nu_{\mathrm{est}}\left(\nu_{\mathrm{est}}+1\right)-2\right]\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*)\mathrm{tr}\left(\mathbf{\Delta}/a\right)}{b\left\{\nu_{\mathrm{est}}^2\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*^2) - \nu_{\mathrm{est}}\mathrm{tr}^2(\widehat{\mathbf{\Sigma}}_*) + 2\left[\nu_{\mathrm{est}}\left(\nu_{\mathrm{est}}+1\right)-2\right]\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*\mathbf{\Delta}/a)\right\}}. \tag{12}$$

**Table 1 Sphericity multipliers for UNIREP power approximations for fixed means**

| | | Multipliers | | | | |
|---|---|---|---|---|---|---|
| **Covariance** | **Test** | $e_1$ | $e_2$ | $e_3$ | $e_4$ | $e_5$ |
| $\mathbf{\Sigma}_*$(known) | Un | 1 | 1 | $\epsilon_n$ | $\epsilon_d$ | $\epsilon_n$ |
| | HF | E $(\widetilde{\epsilon})$ | E $(\widetilde{\epsilon})$ | $\epsilon_n$ | $\epsilon_d$ | $\epsilon_n$ |
| | GG | E $(\widehat{\epsilon})$ | E $(\widehat{\epsilon})$ | $\epsilon_n$ | $\epsilon_d$ | $\epsilon_n$ |
| | Box | $1/b$ | $1/b$ | $\epsilon_n$ | $\epsilon_d$ | $\epsilon_n$ |
| $\widehat{\mathbf{\Sigma}}_*$(estimated) | 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 | $1/b$ | $1/b$ | $\widetilde{\epsilon}_n$ | $\widehat{\epsilon}_d$ | $\widetilde{\epsilon}_n$ |

The corresponding estimator for the null case is $\widetilde{\epsilon}_r = \left[\left(\nu_{\mathrm{est}}+1\right)b\widehat{\epsilon} - 2\right]/\left[b\left(\nu_{\mathrm{est}} - b\widehat{\epsilon}\right)\right]$, the rank-adjusted Huynh-Feldt sphericity estimator.

For estimated covariance and fixed means, approximate estimated UNIREP power is

$$P = 1-F_F\left[F_F^{-1}\left(1-\alpha; e_1\cdot ab, e_2\cdot b\nu_e\right); e_3\cdot ab, e_4\cdot \nu_e b, \frac{\mathrm{tr}\left(\mathbf{\Delta}\right)}{\widehat{\overline{\lambda}}/e_5}\right], \tag{13}$$

with $\widehat{\overline{\lambda}} = \mathrm{tr}(\widehat{\mathbf{\Sigma}}_*)/b$, and $e_1$ through $e_5$ estimated if unknown (Table 1). Nearly every combination of $\widehat{\epsilon}_n$, $\widetilde{\epsilon}_n$, $\widehat{\epsilon}_d$, $\widetilde{\epsilon}_r$, 1 and $1/b$ was examined for each UNIREP test for the wide range of simulations discussed in Muller et al. [5]. The values chosen provided the most accurate results. In retrospect, they are natural choices as well.

### Approximate UNIREP power confidence intervals

The solution to the UNIREP problem parallels the solution to the univariate problem in Taylor and Muller [16]. The methods apply to any general linear hypothesis, including one degree-of-freedom contrasts, such as pair-wise group comparisons and differences in linear trend between two groups. Tests giving scalar secondary parameters are also common for one-group designs and two-group comparisons. For known covariance and means, $e_5$ is defined to be $\epsilon_n$ (Table 1), and the noncentrality in equation 9 is $\omega_* = \left[\mathrm{tr}\left(\mathbf{\Delta}\right)\right]/(\overline{\lambda}/\epsilon_n) = \left[\mathrm{tr}\left(\mathbf{\Delta}\right)\right]/\lambda_{*1}$ with $\overline{\lambda} = \mathrm{tr}(\mathbf{\Sigma}_*)/b$ and $\lambda_{*1} = \overline{\lambda}/\epsilon_n$. For $\epsilon_n = \left[\mathrm{tr}^2(\mathbf{\Sigma}_*) + 2\mathrm{tr}(\mathbf{\Sigma}_*)\mathrm{tr}\left(\mathbf{\Delta}/a\right)\right]\left\{b\left[\mathrm{tr}(\mathbf{\Sigma}_*^2) + 2\mathrm{tr}(\mathbf{\Sigma}_*\mathbf{\Delta}/a)\right]\right\}$. Therefore, it follows that

$$\omega_* = \mathrm{tr}\left(\mathbf{\Delta}\right)\cdot\frac{\mathrm{tr}\left(\mathbf{\Sigma}_*\right) + 2\mathrm{tr}\left(\mathbf{\Delta}/a\right)}{\mathrm{tr}(\mathbf{\Sigma}_*^2) + 2\mathrm{tr}\left(\mathbf{\Delta}\mathbf{\Sigma}_*/a\right)}. \tag{14}$$

For estimated covariance and fixed means, a ratio involving one biased and two unbiased estimators (Lemma B.2) for estimating $\lambda_{*1}$ may be written as

$$\widetilde{\lambda}_{*1} = \frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) + 2\mathrm{tr}(\boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_*/a)}{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*) + 2\mathrm{tr}\,(\boldsymbol{\Delta}/a)}. \tag{15}$$

In parallel to the univariate setting, the distribution of $\widetilde{\lambda}_{*1}$ can be approximated with a Satterthwaite approximation: $\widetilde{\lambda}_{*1}\nu_*/\lambda_{*1} \sim \chi^2(\nu_*)$ with $\nu_* = (b\nu_{\mathrm{est}}) \cdot \widehat{\epsilon}_d/\widetilde{\epsilon}_n$. Lower and upper tail probabilities, $\alpha_L$ and $\alpha_U$, respectively, define the confidence coefficient, $p_{CL} = 1-\alpha_L-\alpha_U$. Also, $c_{\alpha L} = F_{\chi^2}^{-1}(\alpha_L; \nu_*)$ and $c_{\alpha U} = F_{\chi^2}^{-1}(1-\alpha_U; \nu_*)$. Approximate confidence limits for the noncentrality may be calculated using the following:

$$\Pr\left\{c_{\alpha L} < \frac{\widetilde{\lambda}_{*1}\nu_*}{\lambda_{*1}} < c_{\alpha U}\right\} \approx p_{CL} \tag{16}$$

$$\Pr\left\{\frac{c_{\alpha L}}{\widetilde{\lambda}_{*1}\nu_*} < \frac{1}{\lambda_{*1}} < \frac{c_{\alpha U}}{\widetilde{\lambda}_{*1}\nu_*}\right\} \approx p_{CL} \tag{17}$$

$$\Pr\left\{\frac{\mathrm{tr}\,(\boldsymbol{\Delta})\,c_{\alpha L}}{\widetilde{\lambda}_{*1}\nu_*} < \frac{\mathrm{tr}\,(\boldsymbol{\Delta})}{\lambda_{*1}} < \frac{\mathrm{tr}\,(\boldsymbol{\Delta})\,c_{\alpha U}}{\widetilde{\lambda}_{*1}\nu_*}\right\} \approx p_{CL} \tag{18}$$

$$\Pr\left\{\frac{\mathrm{tr}\,(\boldsymbol{\Delta})\,c_{\alpha L}}{\widetilde{\lambda}_{*1}\nu_*} < \omega_* < \frac{\mathrm{tr}\,(\boldsymbol{\Delta})\,c_{\alpha U}}{\widetilde{\lambda}_{*1}\nu_*}\right\} \approx p_{CL}\,. \tag{19}$$

Approximate lower and upper bounds are therefore $\widetilde{\omega}_{*L} = \mathrm{tr}\,(\boldsymbol{\Delta})\,c_{\alpha L}/\widetilde{\lambda}_{*1}\nu_*$ and $\widetilde{\omega}_{*U}\mathrm{tr}\,(\boldsymbol{\Delta})\,c_{\alpha U}/\widetilde{\lambda}_{*1}\nu_*$. The strict monotone dependence of the noncentral $F$ function on the noncentrality ensures an approximate confidence interval for power. Lower and upper bounds on power are, with $e_1$ through $e_4$ defined in Table 1 for $\widehat{\boldsymbol{\Sigma}}_*$,

$$\widetilde{P}_L = 1 - F_F\left[F_F^{-1}(1-\alpha; e_1\cdot ab, e_2\cdot b\nu_e); e_3\cdot ab, e_4\cdot \nu_e b, \widetilde{\omega}_{*L}\right] \tag{20}$$

and

$$\widetilde{P}_U = 1 - F_F\left[F_F^{-1}(1-\alpha; e_1\cdot ab, e_2\cdot b\nu_e); e_3\cdot ab, e_4\cdot \nu_e b, \widetilde{\omega}_{*U}\right]. \tag{21}$$

Taylor and Muller [16] recommended one-sided power confidence intervals by noting 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 [20] provided examples of a one-sided power confidence interval in the univariate case.

### Approximate UNRIEP power confidence regions for power curves

The new methods allow calculating a confidence interval for a single power value. The logic of a proof in Taylor and Muller ([16] equations 2.1-2.13 and surrounding text) guarantees that accurate confidence regions are provided by the point-wise calculations. The proof may be sketched for the present setting as follows. Equations 14-21 establish the validity of the approximate confidence interval for a particular alternative hypothesis, as specified by the scalar constant tr($\boldsymbol{\Delta}$). The randomness in the noncentrality arises from a scalar random variable, $\widetilde{\lambda}_{*1}$, analogous to a variance. Equation 19 describes a single event with a specified probability. The inequality defining the event, and the associated probability, do not change for different values of the scalar constant tr($\boldsymbol{\Delta}$). The smooth and strictly monotone dependence of power on the noncentrality ensures the validity of equations 20-21. The proof is completed by noting that the monotonicity extends the simultaneity property to the power confidence region.

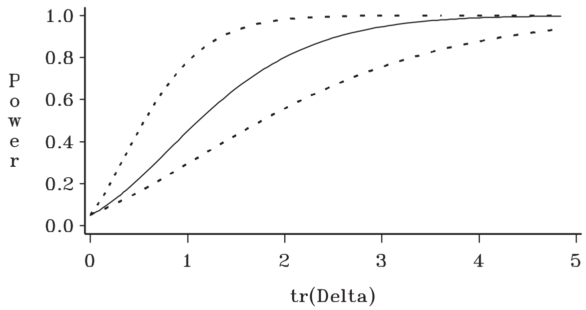

**Figure 1 Approximate 95% confidence region for predicted power of the rank-adjusted Huynh-Feldt test of interaction over tr ($\boldsymbol{\Delta}$) with $N = 10$ and population $\epsilon = 0.720$ for conditions described in Section 'Simulation 1 with rank ($X$) = 1 (one-group repeated measures ANOVA)'.**

Figure 1 gives an example plot of approximate power confidence regions surrounding the predicted power curve for the rank-adjusted Huynh-Feldt test for $\epsilon = 0.720$. Graphical representations such as Figure 1 help researchers accurately recognize the amount of uncertainty in their power calculation, and lead to better decisions about design.

In some cases scientists prefer to consider sample size as a function of the pattern of mean differences. The theory already presented allows plotting sample size as a function of mean difference, albeit with a shift in algorithm. The power function must be numerically inverted to solve for the sample size desired. Taylor and Muller [16] outlined the steps of algorithm needed for the univariate case. Details are not presented here for the sake of brevity.

## Results

### Simulation overview

The accuracy of the new approximate confidence intervals is evaluated for a wide range of conditions. Appendix C contains more details of the simulations and examples. All simulations were conducted in SAS/IML (SAS 9.1, SAS

Institute, 2003) using a version of LINMOD 3.4 modified to include the rank-adjusted Huynh-Feldt estimator and test. Predicted power values and approximate power confidence intervals were computed using a similarly modified version of POWERLIB 2.03. The modified versions of LINMOD and POWERLIB are available at *http://www.health-outcomes-policy.ufl.edu/muller/*.

### Simulation 1 with rank $(X) = 1$ (one-group repeated measures ANOVA)

The accuracy of the new approximate confidence intervals were evaluated for a completely within-subject design with $p = 9$ repeated measures, $N \in \{10, 20, 40\}$, and $q = \text{rank}(X) = 1$. Values for $\boldsymbol{B}$, contrast matrices $\boldsymbol{C}$, $\boldsymbol{U}$, and $\boldsymbol{\Theta}_0$ were chosen to test a within-subject interaction for $\alpha = 0.05$. The model was chosen to ensure predicted power values for the Geisser-Greenhouse test of 0.20, 0.50, and 0.80, using the power approximation in Muller et al. [5]. Population covariance matrices were chosen to provide $\epsilon \in \{0.282, 0.505, 0.720, 1.00\}$. The sphericity values were selected to cover a range of eigenvalue patterns (i.e., patterns of the principal component variances) arising from the structure of $\boldsymbol{\Sigma}_*$. For example, if $b = 3$, then $\boldsymbol{\lambda} = \begin{bmatrix}1 & 0.12 & 0.12\end{bmatrix}'$ gives $\epsilon \approx 0.50$. In turn, $\boldsymbol{\lambda} = \begin{bmatrix}1 & 0 & 0\end{bmatrix}'$ gives $\epsilon = 1/3 \approx 0.33$. Pseudo-random realizations of the error matrix, $\boldsymbol{E}$, were generated and tests were calculated. The observed mean power values for the four UNIREP tests were calculated and tabulated for 500,000 replications per condition.

For the conditions described above, additional pseudo-random realizations of the error matrix were generated using an estimating study with sample size, $N_{\text{est}}$, 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. Both one- and two-sided confidence intervals were evaluated with target coverages of 90% and 95%. The number of replications gave a standard error of estimated coverage probability less than or equal to 0.0003 for $1 - \alpha = 0.95$, and 0.0004 for $1 - \alpha = 0.90$, 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.

**Table 2 Target 95% CI (two-sided) estimated coverage ($\times 100$) of simulated population power for $N = 10$**

| Test | $\epsilon$ | Power | Lower tail | Coverage | Upper tail |
|---|---|---|---|---|---|
| Box | 0.282 | 0.123 | 1.1 | 97.8 | 1.1 |
| | | 0.535 | 1.9 | 97.0 | 1.1 |
| | | 0.930 | 1.7 | 97.3 | 1.0 |
| | 0.505 | 0.054 | 0.1 | 97.3 | 2.6 |
| | | 0.266 | 0.5 | 97.0 | 2.5 |
| | | 0.690 | 1.1 | 97.0 | 1.9 |
| | 0.720 | 0.052 | 0.4 | 94.1 | 5.5 |
| | | 0.227 | 0.6 | 96.8 | 2.6 |
| | | 0.569 | 1.4 | 97.0 | 1.6 |
| | 1 | 0.023 | 0.6 | 85.1 | 14.3 |
| | | 0.117 | 0.5 | 96.0 | 3.5 |
| | | 0.350 | 0.8 | 97.8 | 1.4 |
| GG | 0.282 | 0.155 | 3.1 | 94.7 | 2.2 |
| | | 0.585 | 2.6 | 95.6 | 1.8 |
| | | 0.942 | 1.8 | 96.6 | 1.6 |
| | 0.505 | 0.162 | 5.4 | 87.7 | 6.9 |
| | | 0.520 | 3.8 | 90.6 | 5.6 |
| | | 0.870 | 2.6 | 92.4 | 5.0 |
| | 0.720 | 0.203 | 2.4 | 92.3 | 5.3 |
| | | 0.539 | 2.6 | 94.1 | 3.3 |
| | | 0.856 | 3.3 | 94.2 | 2.5 |
| | 1 | 0.161 | 0.7 | 95.6 | 3.7 |
| | | 0.438 | 1.4 | 97.0 | 1.6 |
| | | 0.751 | 2.7 | 96.2 | 1.1 |
| HF | 0.282 | 0.166 | 3.8 | 93.5 | 2.7 |
| | | 0.602 | 2.8 | 95.2 | 2.0 |
| | | 0.946 | 1.9 | 96.3 | 1.8 |
| | 0.505 | 0.210 | 8.2 | 82.9 | 8.9 |
| | | 0.592 | 4.7 | 88.5 | 6.8 |
| | | 0.902 | 2.9 | 90.9 | 6.2 |
| | 0.720 | 0.271 | 3.6 | 90.9 | 5.5 |
| | | 0.631 | 3.4 | 93.3 | 3.3 |
| | | 0.904 | 4.0 | 93.6 | 2.4 |
| | 1 | 0.224 | 0.8 | 96.7 | 2.5 |
| | | 0.531 | 1.8 | 97.1 | 1.1 |
| | | 0.821 | 3.2 | 95.9 | 0.9 |

Standard error of coverage probability $\times$ 100 $\approx$ 0.0003 $\times$ 100.

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. Table 2 contains results for the Box conservative test with a target sample size of 10. For a wide range of sphericity values and target power values, the target 95% estimated coverage is consistently reached. The two cases in which the target coverage is not reached occur with large population sphericity and low power. Under these conditions, the Box conservative test would not be used in practice.

Table 2 also contains coverage results for the Geisser-Greenhouse and the rank-adjusted Huynh-Feldt tests. The target 95% estimated coverage is consistently reached for extreme sphericity values for both tests. For midrange sphericity values, the coverage fell below the target coverage from 0.8% to 7.3% for the Geisser-Greenhouse, and 1.4% to 12.1% for the rank-adjusted Huynh-Feldt. Coverage accuracy improved as the estimated power increased. In practice, lower power values are of little concern. For target power of 0.80 for the Geisser-Greenhouse test, the largest deviation from the target 95% estimated coverage was 2.6% for the Geisser-Greenhouse test and 4.1% for the rank-adjusted Huynh-Feldt test. Both occurred for the population sphericity value of 0.505.

Only a spherical case is appropriate to consider for the uncorrected test because otherwise the test will have inflated test size. Simulation results in Table 3 show that the approximation for the uncorrected test (with sphericity) always reached the target estimated coverage for the uncorrected test. The conservative bias could be eliminated by using optimal maximum likelihood estimates for the common variance and covariance (Morrison [21]), rather than the unstructured covariance estimate. Additional small changes are needed, associated with degrees of freedom, and corresponding to making all choices of $e_1$ through $e_5$ equal to 1.

Although not presented here, in general, the accuracy of the coverage improved directly with increasing sample size, for all tests and conditions. The accuracy of the approximate confidence bounds for all four UNIREP tests also improved as the population sphericity increased.

### Simulation 2 with rank$(X) > 1$

All of the simulations in the second example considered the condition of rank of $\boldsymbol{X}$ greater than 1. The cases used $p = 5$ repeated measures, $N \in \{16, 32, 48\}$, and $q = \text{rank}\,(\boldsymbol{X}) \in \{2, 4, 8, 16\}$, corresponding to a three-, five-, nine-, and seventeen-group comparison, respectively. 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, $\alpha$, 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. 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 'Simulation 1 with rank $(\boldsymbol{X}) = \mathbf{1}$ (one-group repeated measures ANOVA)'.

**Table 3 Target 95% CI (Two-sided) estimated coverage ($\times$100) of simulated population power for the uncorrected test ($\epsilon = 1.00$)**

| *N* | Population power | Lower tail | Coverage | Upper tail |
|---|---|---|---|---|
| 10 | 0.238 | 0.5 | 97.5 | 2.0 |
|  | 0.551 | 1.5 | 97.6 | 0.9 |
|  | 0.835 | 3.2 | 96.1 | 0.7 |
| 20 | 0.215 | 0.8 | 97.3 | 1.9 |
|  | 0.520 | 1.6 | 97.6 | 0.8 |
|  | 0.814 | 3.1 | 96.2 | 0.7 |
| 40 | 0.207 | 0.9 | 97.2 | 1.9 |
|  | 0.509 | 1.5 | 97.7 | 0.8 |
|  | 0.806 | 3.0 | 96.3 | 0.7 |

Standard error of coverage probability$\times$100$\approx$0.0003$\times$100.

Pseudo-random realizations of the error matrix were generated using an estimating study with sample size, $N_{\text{est}}$, 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 'Approximate UNIREP power confidence intervals'. 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, 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 or 48. 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 5, the proportion of simulations in which the estimated confidence interval successfully covered the observed mean population power values for each test is

**Table 4 Simulated population power for target power = 0.80, $N = 16$ and rank $(X) = q$**

| N | | Simulated population power | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | $\epsilon = 0.282$ | | | $\epsilon = 0.505$ | | | |
| | $q$ | Box | GG | HF | Box | GG | HF | |
| 16 | 2 | 0.779 | 0.811 | 0.817 | 0.561 | 0.778 | 0.809 | |
| | 4 | 0.763 | 0.797 | 0.805 | 0.510 | 0.762 | 0.802 | |
| | 8 | 0.753 | 0.787 | 0.799 | 0.455 | 0.736 | 0.796 | |
| | | $\epsilon = 0.720$ | | | $\epsilon = 1.000$ | | | |
| | $q$ | Box | GG | HF | Box | GG | HF | UN |
| 16 | 2 | 0.457 | 0.760 | 0.805 | 0.399 | 0.748 | 0.790 | 0.799 |
| | 4 | 0.355 | 0.740 | 0.801 | 0.255 | 0.724 | 0.787 | 0.801 |
| | 8 | 0.267 | 0.695 | 0.795 | 0.138 | 0.655 | 0.775 | 0.800 |
| | | $\epsilon = 0.282$ | | | $\epsilon = 0.505$ | | | |
| | $q$ | Box | GG | HF | Box | GG | HF | |
| 48 | 2 | 0.803 | 0.843 | 0.845 | 0.586 | 0.802 | 0.813 | |
| | 4 | 0.773 | 0.812 | 0.815 | 0.552 | 0.794 | 0.805 | |
| | 8 | 0.766 | 0.800 | 0.803 | 0.544 | 0.787 | 0.799 | |
| | 16 | 0.762 | 0.796 | 0.799 | 0.522 | 0.780 | 0.795 | |
| | | $\epsilon = 0.720$ | | | $\epsilon = 1.000$ | | | |
| | $q$ | Box | GG | HF | Box | GG | HF | UN |
| 48 | 2 | 0.500 | 0.792 | 0.806 | 0.455 | 0.785 | 0.797 | 0.800 |
| | 4 | 0.427 | 0.787 | 0.802 | 0.346 | 0.781 | 0.796 | 0.800 |
| | 8 | 0.402 | 0.781 | 0.799 | 0.280 | 0.778 | 0.797 | 0.801 |
| | 16 | 0.359 | 0.775 | 0.799 | 0.221 | 0.770 | 0.795 | 0.801 |

Standard error of estimated power ≈0.0006.

shown. The results are based on using an estimating study with sample size, $N_{\text{est}}$, 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 rank-adjusted 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

**Table 5 Target 95% CI (two-sided) estimated coverage $(\times 100)$ of simulated population power for target power = 0.80, $N = 16$, rank $(X) = q$, $\text{N}_{\text{est}} = 16$, and rank $(X_{\text{est}}) = 4$**

| | $\epsilon = 0.282$ | | | $\epsilon = 0.505$ | | | |
|---|---|---|---|---|---|---|---|
| $q$ | 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 | |
| | $\epsilon = 0.720$ | | | $\epsilon = 1.000$ | | | |
| | 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 |

Standard error of coverage probability × 100 ≈ 0.0003 × 100.

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 rank-adjusted 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 rank-adjusted 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.

The estimated coverages of these tabulated observed mean power values for each test were simulated for population sphericity values of 0.282, 0.505, 0.720, and 1.00. In Table 6, the worst case results from these simulations, which occurred for population sphericity 0.505, are presented. Approximate confidence intervals were simulated for 500, 000 replications per condition (standard error of estimated coverage probability less than or equal to 0.0003 for $1 - \alpha = 0.95$, and 0.0004 for $1 - \alpha = 0.90$). The estimating studies use sample sizes, $N_{\text{est}}$, 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 $\nu_{\text{est}}$ 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 $N_{\text{est}}$ 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 $\nu_{\text{est}}$ 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 $\nu_{\text{est}}$ 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 $N_{\text{est}}$ decreased.

The slow convergence of the approximate power confidence interval coverage to the target coverage may be

**Table 6 Target 95% CI (two-sided) estimated coverage $(\times 100)$ of simulated population power for $\epsilon = 0.505$, target power $= 0.80$, $N = 48$ and rank $(X) = q$**

| | | Box coverage | | | GG coverage | | | HF coverage | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | rank($\mathbf{X}_{\text{est}}$) | | | rank($X_{\text{est}}$) | | | rank($X_{\text{est}}$) | | |
| $N_{\text{est}}$ | $q$ | 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 |

Estimation Study: $N_{\text{est}}$ ∈ (16, 32, 48) and rank ($X_{\text{est}}$ ∈) (2,4,8). Standard error of coverage probability $\times$ 100 $\approx$ 0.0003 $\times$ 100.

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 non-null 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.

### Alternate approximations

In attempts to develop even better confidence bound estimates, additional approximations were developed and evaluated. One approach approximated the distribution of $\widetilde{\lambda}_{*1}$ with an *F*. Using the methods presented in Kim et al. [22], 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 $2\mathrm{tr}\left(\mathbf{\Delta}/a\right)$ component makes the denominator more of a shifted central quadratic. Second, the Kim et al. [22] result requires that the components of the numerator and denominator be mutually independent, which does not hold. Simulations demonstrated that the approximation was inaccurate in small samples.

Alternative approximations were developed and evaluated. The alternatives matched only the numerator to a weighted noncentral chi-square or to a weighted central chi-square with the denominator a constant equal to $\mathrm{E}[\,\mathrm{tr}(\widehat{\mathbf{\Sigma}}_*) + 2\mathrm{tr}(\mathbf{\Delta}/a)]$. All were less accurate than the approximation presented here.

## Discussion

Even for small sample sizes, the proposed power confidence intervals attain very accurate coverage probabilities for the Box conservative test in all cases and for the uncorrected test with $\epsilon = 1$ (the only case for which it should be used). The result is also true for the extreme population sphericity values for the Geisser-Greenhouse and rank-adjusted Huynh-Feldt tests. For midrange population sphericity values, the coverage probabilities of the approximate power confidence intervals for the latter two tests often fell somewhat short of the various target coverage values considered. Coverage probabilities improve as sample size increases. Accuracy is better for higher target power values than for lower, which makes the results useful in practice. One-sided confidence intervals are recommended for lower bounds on power.

The techniques also allow plotting power confidence regions around an estimated power curve (Figure 1). The resulting plots have been well received by researchers.

## Conclusions

Good statistical practice requires associating a credible measure of uncertainty with any parameter estimate. We described and evaluated new methods to meet the need for UNIREP power estimates based on an estimated covariance with fixed means. Across a large range of conditions, the methods provide reasonably accurate coverage for all four UNIREP tests.

## Appendix

### Appendix A: Additional notation

The additional notation comes from Muller et al. [5], who showed that if $S_{t1} = \sum_{k=1}^{b} \lambda_k$, $S_{t2} = \sum_{k=1}^{b} \lambda_k^2$, $S_{t3} = \sum_{k=1}^{b} \lambda_k \omega_k$ and $S_{t4} = \sum_{k=1}^{b} \lambda_k^2 \omega_k$, then

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

$$\nu_{*1} = aS_{t1}/\lambda_{*1} \tag{A.2}$$

$$\omega_* = S_{t3}/\lambda_{*1} \tag{A.3}$$

$$\lambda_{*2} = S_{t2}/S_{t1} \tag{A.4}$$

$$\nu_{*2} = \nu_e S_{t1}^2/S_{t2} = \nu_e b\epsilon. \tag{A.5}$$

They used the parameters (assumed known) to approximate the UNIREP test statistic with a noncentral *F* distribution, as presented in equation 6.

### Appendix B: Supporting lemmas and proofs

**Lemma B.1.** A.1-A.5 imply $\{S_{t1}, S_{t2}, S_{t3}, S_{t4}\} = \{\mathrm{tr}(\mathbf{\Sigma}_*), \mathrm{tr}(\mathbf{\Sigma}_*^2), \mathrm{tr}(\mathbf{\Delta}), \mathrm{tr}(\mathbf{\Sigma}_*\mathbf{\Delta})\}$.

*Proof of lemma B.1.*

$$S_{t1} = \sum_{k=1}^{b} \lambda_k = \sum_{k=1}^{b} \mathrm{tr}\left(\lambda_k \boldsymbol{v}_k' \boldsymbol{v}_k\right) = \sum_{k=1}^{b} \mathrm{tr}\left(\lambda_k \boldsymbol{v}_k \boldsymbol{v}_k'\right)$$

$$= \mathrm{tr}\left[\sum_{k=1}^{b} \left(\lambda_k \boldsymbol{v}_k \boldsymbol{v}_k'\right)\right] = \mathrm{tr}\left(\mathbf{\Sigma}_*\right)$$

$$S_{t2} = \sum_{k=1}^{b} \lambda_k^2 = \sum_{k=1}^{b} \mathrm{tr}\left(\lambda_k^2 \boldsymbol{v}_k' \boldsymbol{v}_k\right) = \sum_{k=1}^{b} \mathrm{tr}\left(\lambda_k^2 \boldsymbol{v}_k \boldsymbol{v}_k'\right)$$

$$= \mathrm{tr}\left[\sum_{k=1}^{b} \left(\lambda_k^2 \boldsymbol{v}_k \boldsymbol{v}_k'\right)\right] = \mathrm{tr}\left(\mathbf{\Sigma}_*^2\right)$$

$$S_{t3} = \sum_{k=1}^{b} \mathrm{tr}\left(\lambda_k \frac{\boldsymbol{v}_k' \boldsymbol{\Delta} \boldsymbol{v}_k}{\lambda_k}\right) = \sum_{k=1}^{b} \mathrm{tr}\left(\boldsymbol{\Delta} \boldsymbol{v}_k \boldsymbol{v}_k'\right)$$

$$= \mathrm{tr}\left[\boldsymbol{\Delta} \sum_{k=1}^{b} \left(\boldsymbol{v}_k \boldsymbol{v}_k'\right)\right] = \mathrm{tr}\left(\boldsymbol{\Delta} \boldsymbol{\Upsilon} \boldsymbol{\Upsilon}'\right) = \mathrm{tr}\left(\boldsymbol{\Delta}\right)$$

$$S_{t4} = \sum_{k=1}^{b} \mathrm{tr}\left(\lambda_k^2 \frac{\boldsymbol{v}_k' \boldsymbol{\Delta} \boldsymbol{v}_k}{\lambda_k}\right) = \sum_{k=1}^{b} \mathrm{tr}\left(\boldsymbol{\Delta} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k'\right)$$

$$= \mathrm{tr}\left[\boldsymbol{\Delta} \sum_{k=1}^{b} \left(\lambda_k \boldsymbol{v}_k \boldsymbol{v}_k'\right)\right] = \mathrm{tr}\left(\boldsymbol{\Delta \Sigma}_*\right).$$

**Lemma B.2.** The first moments of $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)$, $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)$, $\mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)$, and $\mathrm{tr}(\boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_*)$ are known.

*Proof of lemma B.2.* Following Wishart [23], $\boldsymbol{S} = \nu_e \widehat{\boldsymbol{\Sigma}}_* \sim \mathcal{W}_b\left(\nu_e, \boldsymbol{\Sigma}_*\right)$, such that $\nu_e = N - \mathrm{rank}\left(\boldsymbol{X}\right)$. For Wishart $\langle \boldsymbol{\Sigma} \rangle_{jj} = \sigma_j^2$ while here $\langle \boldsymbol{\Sigma} \rangle_{jj} = \sigma_{jj}$ and $\rho_{jk} = \sigma_{jk} / \left(\sigma_{jj}\sigma_{kk}\right)^{1/2}$. Wishart [23] said, with emphasis not in the original, "...moment coefficients are in all cases *except the first* calculated about the mean of the sample...". Here $\mu(n)$ indicates the expression in equation $n$ at the end of Wishart [23], $\mathrm{E}\left[\mathrm{tr}\left(\boldsymbol{S}\right)\right] = \mathrm{E}(\sum_{j=1^b} s_{jj}) = \sum_{j=1}^{b} \mathrm{E}\left(s_{jj}\right) = \sum_{j=1}^{b} \mu(1)_j = \sum_{j=1}^{b} \nu_e \sigma_{jj}$. Thus

$$\mathrm{E}[\,\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)] = (1/\nu_e)\, \mathrm{E}\left[\mathrm{tr}\left(\boldsymbol{S}\right)\right] = (1/\nu_e) \sum_{j=1}^{b} \nu_e \sigma_{jj} = \mathrm{tr}\left(\boldsymbol{\Sigma}_*\right). \tag{B.1}$$

With $\boldsymbol{S}^2 = (\nu_e \widehat{\boldsymbol{\Sigma}}_*)^2$, $\mathrm{E}\left[\mathrm{tr}(\boldsymbol{S}^2)\right] = \mathrm{E}\left[\left(\sum_{j=1}^{b} \sum_{k=1}^{b} s_{jk}^2\right)\right] = \sum_{j=1}^{b} \sum_{k=1}^{b} \mathrm{E}\left(s_{jk}^2\right)$. In turn $\mathrm{E}\left(s_{jk}^2\right) = \mathrm{E}\{[[s_{jk} - \mathrm{E}\left(s_{jk}\right)] + \mathrm{E}\left(s_{jk}\right)]^2\} = \mathrm{E}\left\{\left[s_{jk} - \mathrm{E}\left(s_{jk}\right)\right]^2 + 2\mathrm{E}\left(s_{jk}\right)\left[s_{jk} - \mathrm{E}\left(s_{jk}\right)\right] + \left[\mathrm{E}\left(s_{jk}\right)\right]^2\right\} = \mathrm{E}\left\{\left[s_{jk} - \mathrm{E}\left(s_{jk}\right)\right]^2\right\} + 2\mathrm{E}\left(s_{jk}\right)\mathrm{E}\left[s_{jk} - \mathrm{E}\left(s_{jk}\right)\right] + \left[\mathrm{E}\left(s_{jk}\right)\right]^2 = \mathrm{E}\left\{\left[s_{jk} - \mathrm{E}\left(s_{jk}\right)\right]^2\right\} + \left[\mathrm{E}\left(s_{jk}\right)\right]^2$. Also:

$$\mathrm{E}\left(s_{jk}^2\right) = \begin{cases} \left[s_{jj}^2 - \left[\mathrm{E}\left(s_{jj}\right)\right]^2\right] + \left[\mathrm{E}\left(s_{jj}\right)\right]^2 = \mu(3)_j + \left[\mu\left(1\right)_j\right]^2 = \nu_e \sigma_{jj}^2 \left(2 + \nu_e\right) & \text{if } j = k \\ \left[s_{jk}^2 - \left[\mathrm{E}\left(s_{jk}\right)\right]^2\right] + \left[\mathrm{E}\left(s_{jk}\right)\right]^2 = \mu(5)_{jk} + \left[\mu(2)_{jk}\right]^2 = \nu_e \left(\sigma_{jj}\sigma_{kk} + \sigma_{jk}^2\right) & \text{if } j \neq k. \end{cases}$$

Hence

$$\mathrm{E}[\,\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)] = \left(1/\nu_e^2\right)\left[\nu_e\left(\nu_e + 1\right)\mathrm{tr}(\boldsymbol{\Sigma}_*^2) + \nu_e \mathrm{tr}^2\left(\boldsymbol{\Sigma}_*\right)\right]. \tag{B.2}$$

Here $\mathrm{E}\left[\mathrm{tr}^2\left(\boldsymbol{S}\right)\right] = \mathrm{E}\left[\mathrm{tr}\left(\boldsymbol{S}\right)\mathrm{tr}\left(\boldsymbol{S}\right)\right] = \mathrm{E}\left[\left(\sum_{j=1}^{b} s_{jj}\right)\left(\sum_{k=1}^{b} s_{kk}\right)\right] = \mathrm{E}\left(\sum_{j=1}^{b}\sum_{k=1}^{b} s_{jj}s_{kk}\right) = \sum_{j=1}^{b}\sum_{k=1}^{b} \mathrm{E}\left(s_{jj}s_{kk}\right)$, with $\mathrm{E}\left[s_{jj}s_{kk}\right] = \mu(4)_{jk} = 2\nu_e\sigma_{jk}^2 + \nu_e^2\sigma_{jj}\sigma_{kk}$. Thus,

$$\begin{aligned} \mathrm{E}\left[\mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)\right] &= \left(1/\nu_e^2\right)\mathrm{E}\left[\mathrm{tr}^2\left(\boldsymbol{S}\right)\right] \\ &= \left(1/\nu_e^2\right)\sum_{j=1}^{b}\sum_{k=1}^{b}\left(2\nu_e\sigma_{jk}^2 + \nu_e^2\sigma_{jj}\sigma_{kk}\right) \\ &= \left(1/\nu_e^2\right)\left[\,2\nu_e\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) + \nu_e^2\mathrm{tr}^2\left(\boldsymbol{\Sigma}_*\right)\,\right]. \end{aligned} \tag{B.3}$$

Finally, $\boldsymbol{S} = \nu_e\widehat{\boldsymbol{\Sigma}}_* \sim \mathcal{W}_b\left(\nu_e, \boldsymbol{\Sigma}_*\right)$ has $\mathrm{E}(\boldsymbol{S}) = \nu_e\boldsymbol{\Sigma}_*$ (Muller and Stewart [6], Theorem 10.10). Hence $\mathrm{E}[\mathrm{tr}(\boldsymbol{\Delta S})] = \mathrm{tr}\left[\mathrm{E}(\boldsymbol{\Delta S})\right] = \mathrm{tr}\left[\boldsymbol{\Delta}\mathrm{E}(\boldsymbol{S})\right] = \mathrm{tr}\left[\boldsymbol{\Delta}(\nu_e\boldsymbol{\Sigma}_*)\right] = \nu_e\mathrm{tr}(\boldsymbol{\Delta\Sigma}_*)$ and

$$\mathrm{E}\left[\mathrm{tr}(\boldsymbol{\Delta}\widehat{\boldsymbol{\Sigma}}_*)\right] = \mathrm{tr}(\boldsymbol{\Delta\Sigma}_*). \tag{B.4}$$

*Proof of lemma 1.* Substituting equivalent terms from equations A.1-A.5 into $(\lambda_{*2}ab\nu_{*2}) / (\lambda_{*1}\nu_{*1}b\nu_e)$ allows combining like terms and simplifying to give

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

If $T_u = [\,\mathrm{tr}(\widehat{\boldsymbol{\Delta}})/a] / [\,\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)]$, then $\mathrm{tr}(\widehat{\boldsymbol{\Delta}}) \approx \lambda_{*1}y_{*1}$ and $\nu_e\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*) \approx \lambda_{*2}y_{*2}$ with $y_{*1} \sim \chi^2\left(\nu_{*1}, \omega_*\right)$ and $y_{*2} \sim \chi^2\left(\nu_{*2}\right)$ as described in Muller et al. [5]. In turn, $\Pr\left\{T_u \leq t\right\} = \Pr\left\{[\,\mathrm{tr}(\widehat{\boldsymbol{\Delta}})/a] / \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*) \leq t\right\} \approx \Pr\left\{\left[\lambda_{*1}y_{*1}/(ab)\right] / \left[\lambda_{*2}y_{*2}/\left(b\nu_e\right)\right] \leq t\right\} = \Pr\left\{\left(y_{*1}/\nu_{*1}\right) / \left(y_{*2}/\nu_{*2}\right) \leq t\lambda_{*2}ab\nu_{*2}/\left(\lambda_{*1}\nu_{*1}b\nu_e\right)\right\} = F_F\left(t; \nu_{*1}, \nu_{*2}, \omega_*\right)$.

*Proof of lemma 2.* Using Lemma B.2, unbiased estimators for $\mathrm{tr}^2(\boldsymbol{\Sigma}_*)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_*^2)$ are $\widehat{\tau}_1 = [\,\mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2\left(\nu_e + 1\right)^{-1}\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2)]\,\{1 - 2\left[\nu_e\left(\nu_e + 1\right)\right]^{-1}\}^{-1}$ and $\widehat{\tau}_2 = [\,\nu_e^2\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e\mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*)]\left[\nu_e\left(\nu_e + 1\right) - 2\right]^{-1}$, as introduced in Gribbin [18] and Chi et al. [19]. As shown in Lemma B.2, $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)$ and $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta})$ are unbiased estimators for $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)$ and $\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta})$, respectively. Thus,

$$\widetilde{\epsilon}_n = \frac{\widehat{\tau}_1 + 2\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)\mathrm{tr}\left(\boldsymbol{\Delta}/a\right)}{b\left[\widehat{\tau}_2 + 2\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\right]} = \frac{\nu_e\left(\nu_e+1\right)\mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) - 2\nu_e\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) + 2\left[\nu_e\left(\nu_e+1\right)-2\right]\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*)\mathrm{tr}\left(\boldsymbol{\Delta}/a\right)}{b\left\{\nu_e^2\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*^2) - \nu_e\mathrm{tr}^2(\widehat{\boldsymbol{\Sigma}}_*) + 2\left[\nu_e\left(\nu_e+1\right)-2\right]\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_*\boldsymbol{\Delta}/a)\right\}}.$$

In the null case $\boldsymbol{\Delta} = \mathbf{0}$ and $\widetilde{\epsilon}_n$ reduces to the rank-adjusted sphericity estimator, $\widetilde{\epsilon}_r = [(\nu_e+1)\, b\widehat{\epsilon} - 2] / [b\,(\nu_e - b\widehat{\epsilon})] = \widetilde{\epsilon}_n | \boldsymbol{\Delta} = 0$.

## Appendix C: Simulation details

### Covariance conditions

Covariance conditions 5-8 from Table III of Coffey and Muller [24] were used for each example described below: $\boldsymbol{\Sigma}_* = \mathrm{Dg}\left(\boldsymbol{\lambda}_j\right)$ for $j \in \{1, 2, 3, 4\}$, with

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

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

### CLAHE mammography example

Computer scientists developed the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm to improve contrast in medical images. Independent observers considered $3 \times 3 = 9$ Clip×Region combinations. 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 $\boldsymbol{B}$, $(1 \times 9)$, contained mean $\log_{10}$ (contrast) for the unprocessed condition minus the mean for each of the nine combinations of Clip and Region ($\beta_{\mathrm{cr}} = \mu_{\mathrm{unprocessed}} - \mu_{\mathrm{cr}}$). If $\boldsymbol{T}_{\mathrm{c}}$ contains orthonormal linear and quadratic trends for $\log_2$ (Clip) $\in \{1, 2, 4\}$, and $\boldsymbol{T}_{\mathrm{r}}$ does the same for $\log_2\left(\mathrm{Region}\right) \in \{1, 3, 5\}$, then the $9 \times 4$ within-persons contrast matrix, $\boldsymbol{U}_{\mathrm{cr}}$ is

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

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

All four covariance patterns were factorially combined with $N \in \{10, 20, 40\}$. The multivariate test considered $\boldsymbol{\Theta}_{\mathrm{cr}} = \beta_P \cdot \left[0.5\ 1.0\ -1.0\ 0.5\right]$ with $\alpha = 0.05$, and $\beta_P$ 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 in Muller et al. [5]. The conditions in the example were used in section 'Simulation 1 with rank $(\boldsymbol{X}) = \mathbf{1}$ (one-group repeated measures ANOVA)'. More details of the example are in Muller et al. [5].

### Test of interaction with rank($X$) > 1 Example

All cases used 5 repeated measures, $N \in \{16, 32, 48\}$, and $\mathrm{rank}(\boldsymbol{X}) \in \{2, 4, 8, 16\}$. 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, $\boldsymbol{X} = \boldsymbol{I}_q \otimes \mathbf{1}_{repn}$, such that $repn = N/q$, and $\otimes$ is a Kronecker product. If

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

then $\boldsymbol{B} = \beta_P \cdot \boldsymbol{D}_a$, such that $\beta_P$ was the scaling factor giving approximate target power $P \in \{0.20, 0.50, 0.80\}$, for the rank-adjusted Huynh-Feldt power approximation. The within-subject contrast, $\boldsymbol{U}$, $(5 \times 4)$, contained orthonormal linear, quadratic, cubic and quartic trends:

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

Each row of the between-subject contrast, $\boldsymbol{C}$, a $(q - 1 \times q)$ orthonormal matrix, contained one of the $(q - 1)$ trends. The contrasts define a test of interaction of between- and within-subject trends. Without loss of generality, we assumed $\boldsymbol{\Theta}_0 = \mathbf{0}$. A test size, $\alpha$, of 0.05 was used.

### Computational methods

All power computations were conducted in SAS/IML (SAS 9.1, SAS Institute, 2003). Free software LINMOD 3.4 was used for all data analysis and includes new methods. Free software POWERLIB 2.1 in Johnson et al. [25] was used for all power analysis and includes the new methods. Both are available at *http://health-outcomes-policy.ufl.edu/faculty-directory/-muller-keith/list-of-software/*. UNIREP power is also available in GLIMMPSE, a free web-browser based program with a graphical user interface aimed at health scientists (*www.SampleSizeShop.org*). The next version of GLIMMPSE is expected to implement the confidence interval methods.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

The majority of the work comes directly from the dissertation of MJG at UNC Chapel Hill, in partial fulfillment of the requirements of the DrPH, under the direction of KEM. Dr. KEM and Dr. Y-YC assisted in revising and compressing the notation, text and tables. Dr. PWS was a member of the dissertation committee and assisted in formulating the proof of simultaneous coverage of the confidence regions and in final editing. All authors read and approved the final manuscript.

### Acknowledgements

We thank the reviewers Drs. Warren Comulada, Peter Lachenbruch, and Russell Lenth for their timely review and beneficial comments. This work was supported in part by a UF CTSI core grant to Chi and Muller (NCATS UL1TR000064). Chi's support included NIDCR R01-DE020832, NIDDK R01-DK072398, NIDA R01-DA031017, and NICHD P01-HD065647. Muller's support included NIDCR R01-DE020832-01A1, NIDDK R01-DK072398, NIDCR U54-DE019261, NCRR K30-RR022258, NHLBI R01-HL091005, NIAAA R01-AA013458-01, and NIDA R01-DA031017.

### Author details

[1]Department of Biostatistics, MedImmune, Gaithersburg, MD, USA.
[2]Department of Biostatistics, University of Florida, Gainesville, FL, USA.
[3]Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA.
[4]Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, USA.

Received: 16 October 2012 Accepted: 15 March 2013
Published: 15 April 2013

### References

1. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage, Jerome A: **Caution regarding the use of pilot studies to guide power calculations for study proposals.** *Arch Gen Psychiatry* 2006, **63:**484–489.
2. Maxwell SE: **The persistence of underpowered studies in psychological research:causes, consequences, and remedies.** *Psychol Methods* 2004, **9:**147–163.
3. Taylor DJ, Muller KE: **Bias in linear model power and sample size calculation due to estimating noncentrality.** *Commun Stat Theory Methods* 1996, **25:**1595–1610.
4. Muller KE, Pasour VB: **Bias in linear model power and sample size due to estimating variance.** *Commun Stat Theory Methods* 1997, **26:**839–851.
5. Muller KE, Edwards LJ, Simpson SL, Taylor DJ: **Statistical tests with accurate size and power for balance linear mixed models.** *Stat Med* 2007, **26:**3639–3660.
6. Muller KE, Stewart PW: *Linear Model, Theory: Univariate, Multivariate and Mixed Models*. New York: Wiley; 2006.
7. Box GEP: **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.** *Ann Math Stat* 1954, **25:**290–302.
8. Box GEP: **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.** *Ann Math Stat* 1954, **25:**484–498.
9. Geisser S, Greenhouse SW: **An extension of Box's results on the use of the $F$ distribution in multivariate analysis.** *Ann Math Stat* 1958, **29:**885–891.
10. Greenhouse SW, Geisser S: **On methods in the analysis of profile data.** *Psychometrika* 1959, **24:**95–112.
11. Huynh H, Feldt LS: **Estimation of the Box corrections for degrees of freedom from sample data in randomized block and split-plot designs.** *J Edu Stat* 1976, **1:**69–82.
12. Davies RB: **Distribution of a linear combination of chi-square random variables.** *Appl Stat* 1980, **29:**323–333.
13. Muller KE, Barton CN: **Approximate power for repeated measures ANOVA lacking sphericity.** *J Am Stat Assoc* 1989, **84:**549–555.
14. Muller KE, Barton CN: **Correction to Approximate power for repeated-measures ANOVA lacking sphericity.** *J Am Stat Assoc* 1991, **86:**255–256.
15. Browne RH: **On the use of a pilot sample for sample size determination.** *Stat Med* 1995, **14:**1933–1940.
16. Taylor DJ, Muller KE: **Computing confidence bounds for power and sample size of the general linear univariate model.** *Am Stat* 1995, **49:**43–47.
17. Lecoutre B: **A correction for the $\widetilde{\epsilon}$ approximate test with repeated measures design with two or more independent groups.** *J Educ Stat* 1991, **16:**371–372.
18. Gribbin MJ: **Better power methods for the univariate approach to repeated measures.** *PhD Dissertation* 2007. University of North Carolina at Chapel Hill, Department of Biostatistics.
19. Chi YY, Gribbin MJ, Lamers Y, Gregory JF, Muller KE: **Global hypothesis testing for high-dimensional repeated measures outcomes.** *Stat Med* 2012, **31:**724–742.
20. Muller KE, Fetterman BA: *Regression and, ANOVA: An Integrated Approach Using SASő Software Chapter 17*. Cary: SAS Institute; 2002.
21. Morrison DF: *Multivariate Statistical, Methods*. 3rd ed. New York: McGraw-Hill; 1990.
22. Kim HY, Gribbin MJ, Muller KE, Taylor DJ: **Analytic, computational and approximate forms for ratios of noncentral and central Gaussian quadratic forms.** *J Comput Graphical Stat* 2006, **15:**443–459.
23. Wishart J: **The generalized product moment distribution in samples from a normal multivariate population.** *Biometrika* 1928, **20A:**32–52.
24. Coffey CS, Muller KE: **Properties of internal pilots with the univariate approach to repeated measures.** *Stat Med* 2003, **22:**2469–2485.
25. Johnson JL, Muller KE, Slaughter JC, Gurka MJ, Gribbin MJ, and Simpson SL: **POWERLIB: SAS/IML software for computing power in multivariate linear models.** *J Stat Softw* 2009, **30:**1–27. [http://www.jstatsoft.org/v30/i05/paper]

doi:10.1186/1471-2288-13-57
**Cite this article as:** Gribbin *et al.*: Confidence regions for repeated measures ANOVA power curves based on estimated covariance. *BMC Medical Research Methodology* 2013 **13**:57.