BOOTSTRAP RESAMPLING APPROACHES FOR REPEATED MEASURE DESIGNS: RELATIVE ROBUSTNESS TO SPHERICITY AND NORMALITY VIOLATIONS

ILONA BERKOVITS AND GREGORY R. HANCOCK
University of Maryland, College Park

JONATHAN NEVITT
University of Maryland School of Medicine

The current article proposes a bootstrap-F method and a bootstrap-T^2 method for use in a one-way repeated measure ANOVA design. Using a Monte Carlo approach in which sample size, nonsphericity, and nonnormality are systematically manipulated, the Type I error rate of the two bootstrap methods are compared to that of the traditional F test, the Geisser-Greenhouse adjusted F test, the Box adjusted F test, the Huynh-Feldt adjusted F test, the β-trimmed mean method using $\beta = .1$ and $\beta = .2$, and the one-sample multivariate T^2 test. Results show the bootstrap-F method controls Type I error better than all other methods considered when normality and sphericity assumptions are violated simultaneously.

When violations of the sphericity assumption occur in designs containing repeated measures, particularly when compounded by nonnormality, the best strategy for analysis is not entirely clear. As the sphericity assumption alone becomes more severely violated, the traditional unadjusted within-subjects F test is known to perform quite poorly, with its Type I error rate becoming extremely inflated (e.g., Lix, Keselman, & Keselman, 1994). Under such circumstances the researcher has a host of methodological alternatives. Multivariate analyses, although not requiring sphericity, are dependent on normality and apparently are quite sensitive to extreme skew (Harwell & Serlin, 1997); they also are not particularly powerful with smaller sample sizes. Among univariate approaches, a variety of alternatives do exist.

Educational and Psychological Measurement, Vol. 60 No. 6, December 2000 877-892

One of these alternatives is a method using trimmed-means and Winsorized (co)variances. However, in the one-way repeated measures ANOVA, the trimmed means method tests a modified null hypothesis pertaining to the equality of the population trimmed means across repeated measures, rather than the equality of the usual population means, which may render their use unacceptable in particular research scenarios. On the other hand, some researchers (cf. Wilcox, 1997, 1998) have argued that researchers should be more interested in trimmed means than in classical statistics because results from these analyses are more robust and may therefore be more accurate and replicable.

In the approach invoking either trimming, β-trimmed means are means calculated from ordered data with the desired β proportion of data removed from both the upper and lower tail of the distribution. Wilcox (1993) explored the effects of nonnormality on the method trimmed-means, finding that "inferences based on the trimmed mean can have substantially more power [than inferences based on the traditional mean]" (p. 75). The current simulation study includes an investigation of the behavior of two levels of β-trimmed means, $\beta = .1$ and $\beta = .2$, under varied levels of nonnormality and nonsphericity.

Another class of methods uses *F* distributions with degree of freedom (*df*) adjustments to combat sphericity violations. In these methods the test associated with an *a*-level repeated measure for *n* cases is conducted using $df_{\text{numerator}} = (a-1)\,\varepsilon$ and $df_{\text{denominator}} = (a-1)(n-1)\,\varepsilon$, where ε would ideally be the population sphericity. Three approximations for this sphericity *df* adjustment commonly offered in statistical software packages are the Geisser-Greenhouse (GG) lower-bound adjustment (Geisser & Greenhouse, 1958), Box's (1954) $\hat{\varepsilon}$ adjustment (which, curiously, the SPSS package refers to as "Greenhouse-Geisser"), and the Huynh-Feldt (HF) $\tilde{\varepsilon}$ *df* adjustment (Huynh & Feldt, 1976). The GG lower-bound adjustment acknowledges that the worst possible population scenario would have $\varepsilon = 1/(a-1)$; thus the observed univariate *F* statistic is compared to a conservative distribution with $df_{\text{numerator}} = 1$ and $df_{\text{denominator}} = n - 1$. Box (1954) proposed the less conservative *df* adjustment $\hat{\varepsilon}$, expressed by Maxwell and Delaney (1990) as

$$\hat{\varepsilon} = \frac{a^2(\overline{E}_{jj} - \overline{E}_{..})^2}{(a-1)[(\Sigma\Sigma E_{jk}^2) - (2a\Sigma\overline{E}_{j.}^2) + (a^2\overline{E}_{..}^2)]},$$

where *a* is the number of levels of the repeated measure factor, E_{jk} is an element in row *j* and column *k* of the sample covariance matrix, $\overline{E}_{jj}$ is the mean of the diagonal entries (variances) in the sample covariance matrix, $\overline{E}_{j.}$ is the mean of the entries in the *j*th row of the sample covariance matrix, and $\overline{E}_{..}$ is the mean of all entries in the sample covariance matrix. Finally, Huynh and

Feldt (1976) proposed $\tilde{\varepsilon}$, a slightly less conservative modification of Box's statistic that is expressed by Maxwell and Delaney (1990) as

$$\tilde{\varepsilon} = \frac{n(a-1)\hat{\varepsilon} - 2}{(a-1)[n-1-(a-1)\hat{\varepsilon}]}.$$

Some authors (e.g., Keppel, 1991; Lomax, 1998) even discuss using these *df*-adjusted methods in a more complex decision structure originally proposed by Greenhouse and Geisser (1959), whereby the null hypothesis is rejected only if the traditional unadjusted *F*, the GG lower bound, and Box's $\hat{\varepsilon}$ yield collectively at least two null rejections. This strategy was designed to minimize the need for computing Box's cumbersome correction, except in the case where the unadjusted *F* statistic rejects the null hypothesis and the GG lower bound does not; only then would Box's $\hat{\varepsilon}$ be needed as a tie breaker. However, given that Box's approach is always more conservative than the unadjusted *F* test and less so than the GG lower-bound correction, the reader may verify logically that this complex decision strategy will always result in the same conclusion as Box's $\hat{\varepsilon}$ alone. With current software packages providing Box's $\hat{\varepsilon}$, then, the need to implement such a strategy becomes moot.

Unfortunately, the *df*-adjusted methods are only designed to withstand the effects of nonsphericity; nonnormality, in addition to sphericity violations, may lead to nonrobustness either in the form of liberalism or conservatism (Tandon & Moeschberger, 1989), depending on the nature of the *df* adjustment. Still, the principle behind the *df*-adjusted methods remains sound. Specifically, if the traditional distribution does not describe the behavior of the test statistic of interest, one may find another distribution that does (in this case, by adjusting the *df*). This statement is reminiscent of the premise behind recent bootstrap resampling approaches attributed to Efron and colleagues (e.g., Diaconis & Efron, 1983; Efron, 1979; Efron & Gong, 1983), whereby an empirical sampling distribution for the test statistic of interest is derived by repeatedly resampling (with replacement) from the sample at hand. These methods have generally been shown to represent a promising new data analysis paradigm, particularly when assumptions underlying traditional statistical methods are violated, or when evaluating statistics (e.g., either trimmed or Winsorized means) for which sampling distributions are not known. Lunneborg's (2000) recent book provides a comprehensive summary of some of these applications.

In the specific context of experimental designs, Westfall and Young (1993) described a host of bootstrapping applications proposed to be more robust than traditional ANOVA methodology. Unfortunately, these authors provided little information regarding bootstrapped omnibus tests in repeated measure designs. In addition, Lunneborg and Tousignant (1985) did propose a method within the context of repeated measure designs; however, their

approach assumed particular design matrices based on the quantitative nature of the repeated measure variable and also was subsequently shown to be liberal in its Type I error control (Rasmussen, 1987).

As described below, the purpose of the current study is (a) to propose bootstrap resampling F and T^2 procedures for the one-way repeated measure design and (b) to evaluate the robustness of these procedures to sphericity and normality violations. Using Monte Carlo methods to create varied nonnormal and nonspherical conditions, the Type I error rates of the bootstrap-F and bootstrap-T^2 method will be compared to the Type I error rates of the traditional one-sample multivariate T^2 test, to the traditional unadjusted F, to the β-trimmed mean method using β = .1 and β = .2, as well as to three *df*-adjusted methods that, like the bootstrap, use alternate distributions to judge an observed test statistic.

Method

Bootstrapping in a One-Way Repeated Measure Design

For a design with n cases repeating all a levels of within-subjects factor A, a score $\overline{Y}_{ij}$ in any of the $n \times a$ cells of the design may be represented as $Y_{ij} = \overline{Y}_{.j} + e_{ij}$. That is, each score may be regarded as a case's individual error within the jth level of factor A added to the mean of that jth level, $\overline{Y}_{.j}$. Thus, the data for the ith case may be expressed as a row vector $[(\overline{Y}_{.1} + e_{i1})\ (\overline{Y}_{.2} + e_{i2})...(\overline{Y}_{.a} + e_{ia})]$. The data within the a levels exhibit particular distributional properties, and the data across all levels have covariance matrix $\boldsymbol{S}$ where

$$\boldsymbol{S} = \left[\frac{1}{n-1}\right]\begin{bmatrix} e_{11} & \cdot & \cdot & e_{1a} \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ e_{n1} & \cdot & \cdot & e_{na} \end{bmatrix}'\begin{bmatrix} e_{11} & \cdot & \cdot & e_{1a} \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ e_{n1} & \cdot & \cdot & e_{na} \end{bmatrix}.$$

The two bootstrap methods proposed in the current study, the bootstrap-F and the bootstrap-T^2, use the data from a given sample, which have distributional and sphericity properties that estimate (and behave as proxy for) those of the population. To create a null condition in the sample data, centering is performed whereby the mean of each jth level of the repeated measure is subtracted from the cases' data within that level. Thus, the data for the ith case may now simply be expressed as a row vector $[e_{i1}\ e_{i2}\ ...\ e_{ia}]$. These centered data within the a levels still exhibit the same distributional properties as the original sample, and the centered data across all levels still have the same covariance matrix $\boldsymbol{S}$. The means of all a levels are now identical (in fact, all

are 0), creating a null parent sample from which to conduct bootstrap resampling.

In general, bootstrapping is conducted by randomly resampling (with replacement) *n*-centered case vectors from the original *n*-centered case vectors, generating a bootstrapped data set. From this bootstrapped data set, omnibus test statistics are computed (an asterisk will indicate computations to be performed on a bootstrapped data set). For the current study, the F^* and T^{2*} statistics are computed for each bootstrapped data set, where $F^* = MS_A^*/MS_{S\times A}^*$, and $T^{2*} = n\,\mathbf{y}_d^{*\prime}\,\mathbf{S}_d^{*-1}\,\mathbf{y}_d^*$ (see Stevens, 1996). This process is repeated *b* times to create empirical sampling distributions of F^* and T^{2*} values, generated under the original sample's distributional and sphericity properties. To determine the *p* value for a bootstrap procedure, the test statistic for the original sample (observed F or T^2) is placed within the corresponding empirical sampling F^* or T^{2*} distribution. The proportion of F^* or T^{2*} values larger than the observed F or T^2, respectively, represents the bootstrap *p* value. This bootstrap *p* value is used to gauge the tenability of the omnibus null hypothesis $\mu_1 = \mu_2 = \ldots = \mu_a$. The Type I error robustness of these two bootstrap methods to violations of sphericity and normality will be compared to the original F and T^2 tests' performances, as well as to popular *df*-adjusted methods using the following Monte Carlo simulation.

Monte Carlo Simulation

A Monte Carlo simulation was designed to investigate Type I error rate robustness under the following factorially crossed conditions. Sphericity levels were chosen as $\varepsilon = 1.00$, $\varepsilon = .75$, $\varepsilon = .57$, and $\varepsilon = .48$, thereby ranging from perfect sphericity to a severe violation that would be considered within "roughly the lower limits of values often reported in the behavioral literature" (Quintana & Maxwell, 1994, p. 62). The population covariance matrices for these levels of sphericity, which were used to guide data generation, are listed in Table 1. Covariance matrices for the nonspherical conditions are the same as in a study by Keselman and Keselman (1990). The population distribution for all of the *a* levels of the repeated measure were set as normal (skew = 0, kurtosis = 0), slightly nonnormal (skew = 1.00, kurtosis = 0.75), moderately nonnormal (skew = 1.75, kurtosis = 3.75), and severely nonnormal (skew = 3.00, kurtosis = 21.00). Sample sizes investigated were $n = 10$, 15, 30, and 60. For this study, the number of levels of the repeated measure was held constant at $a = 4$, and a nominal error rate of $\alpha = .05$ was adopted.

For each of the $4 \times 4 \times 4 = 64$ cells of the design, 10,000 simulated data sets (i.e., replications) were created in the software package GAUSS (Aptech Systems, 1996) using methods based on the work of Fleishman (1978) and Vale and Maurelli (1983); these methods are described in the next section. For each simulated set of sample data, six of the nine tests were evaluated: the

Table 1
Population Covariance Matrices for Target Sphericity Levels

ε = 1.00				ε = 0.75			
10.0	5.0	5.0	5.0	18.0	8.0	6.0	4.0
5.0	10.0	5.0	5.0	8.0	8.0	5.0	4.0
5.0	5.0	10.0	5.0	6.0	5.0	7.0	3.0
5.0	5.0	5.0	10.0	4.0	4.0	3.0	7.0
ε = 0.57				**ε = 0.48**			
23.2	11.8	7.4	2.4	22.3	10.8	6.5	1.9
11.8	10.3	5.3	1.7	10.8	8.3	5.2	3.1
7.4	5.3	4.3	1.4	6.5	5.2	4.7	2.5
2.4	1.7	1.4	2.2	1.9	3.1	2.5	4.7

traditional one-sample T^2, the traditional unadjusted F, the GG lower-bound df-adjusted method, Box's $\hat{\varepsilon}$ approach, the HF $\tilde{\varepsilon}$ corrected method, and the currently proposed bootstrap-F. The decision to investigate the β-trimmed means and the bootstrap-T^2 was made after the previous six methods had been evaluated, and thus different sample data were simulated from identical populations for its investigation. Notwithstanding, the new 10,000 data sets generated per cell for investigation of the β-trimmed means and the bootstrap-T^2 should yield results that may be compared with those of the previous six methods.

For the nonbootstrap tests, p values were found using the appropriate reference distributions as contained within the GAUSS software package (Aptech Systems, 1996). Bootstrap p values for the proposed methods were determined as previously described, based on $b = 5{,}000$ bootstrapped samples of size n from that simulated data set.

A test of the omnibus null hypothesis was performed for each method: If the $p < .05$, a rejection decision was recorded for that method on that replication. It should be noted that the bootstrap-T^{2*} could not be computed for less than 1% of bootstrapped data sets that had singular or near singular covariance matrices; in such cases, to compute the T^{2*} these problematic data sets were replaced by new bootstrapped data sets where the determinant of the covariance matrix was larger than .001. For all methods under all conditions, then, the proportion of Type I errors was determined as the number of false rejections out of 10,000 replications. As is commonly done, Bradley's (1978) liberal criterion was used to assess the robustness of each method under each condition. According to this criterion, the test's empirical Type I error rate ($\hat{\alpha}$) must be contained in the interval $.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$ to be considered robust. Therefore, for the $\alpha = .05$ level of statistical significance used in this study, the interval used to define robustness was $.025 \leq \hat{\alpha} \leq .075$.

Data Generation

Using GAUSS (Aptech Systems, 1996), data were simulated to conform to each of the 16 sphericity and distributional condition combinations. Multivariate normal and nonnormal data were generated via the algorithm developed by Vale and Maurelli (1983), which is a multivariate extension of the method for simulating univariate data proposed by Fleishman (1978). Programming used in this study to generate simulated data has been scrutinized externally and verified for accuracy and is available to the general research community (Nevitt & Hancock, 1999).

The above procedures yielded an $n \times 4$ standardized data matrix whose marginal distributions have univariate skew and kurtosis matching the target distributional form and with the desired correlational structure (i.e., the standardized target covariance matrix). The final step in the data generation process was to convert the correlational structure back to the covariance structure associated with the target level of sphericity. This was accomplished by multiplying the $n \times 4$ data matrix by a 4×4 diagonal matrix of standard deviations, each element being the square root of a diagonal element from a given population covariance matrix shown in Table 1.

As a final verification of the data generation mechanism, one simulated data matrix was drawn from each of the 16 population conditions using a very large sample size (n = 100,000). From each simulated 100,000 $\times$ 4 data matrix, large sample estimates of skew and kurtosis were obtained for each of the four column vectors in the data matrix. Unbiased estimates for skew and kurtosis were computed using the Fisher g statistics (see, e.g., DeCarlo, 1997, p. 301); a sample estimate of ε was also obtained for each of the large sample data matrices. Results from these verification samples are extremely close to the target population parameters, as presented in the appendix, leaving us quite confident that the data generation mechanism yields simulated data that conform closely to both the target distributional form and the target level of sphericity.

Results

The empirical Type I error rates under all conditions for the two bootstrapping approaches and the seven other tests appear in Table 2, whereas Table 3 contains each method's number of cells in the design outside of Bradley's liberal robustness interval. These results in Table 3 are categorized into four types of underlying populations: normal and spherical (4 cells), normal and nonspherical (12 cells), nonnormal and spherical (12 cells), and nonnormal and nonspherical (36 cells). It should be noted that the Type I error rates for each condition were also evaluated separately using a 95% confidence interval around each $\hat{\alpha}$ to determine if $\alpha = .05$ was contained within the confidence interval. Although some methods would be considered nonrobust using this

Table 2
Empirical Type I Error Rates for Each Method Under All Conditions

ε	*F*	HF	Box	GG	.1-Trimmed	.2-Trimmed	T^2	Bootstrap-*F*	Bootstrap-T^2
n = 10									
skew = 0, kurt = 0									
1.0	0.0486	0.0444	0.0348	0.0061^{C}	0.0379	0.0358	0.0473	0.0316	0.0072^{C}
.75	0.0657	0.0543	0.0410	0.0129^{C}	0.0491	0.0436	0.0460	0.0372	0.0070^{C}
.57	0.0802^{L}	0.0560	0.0452	0.0193^{C}	0.0542	0.0527	0.0469	0.0401	0.0070^{C}
.48	0.0893^{L}	0.0567	0.0484	0.0294	0.0552	0.0557	0.0493	0.0436	0.0074^{C}
skew = 1.00, kurt = 0.75									
1.0	0.0495	0.0428	0.0304	0.0054^{C}	0.0344	0.0309	0.0460	0.0281	0.0049^{C}
.75	0.0721	0.0575	0.0444	0.0136^{C}	0.0580	0.0546	0.0564	0.0398	0.0091^{C}
.57	0.0897^{L}	0.0702	0.0606	0.0367	0.0860^{L}	0.0867^{L}	0.0782^{L}	0.0509	0.0130^{C}
.48	0.1056^{L}	0.0718	0.0630	0.0406	0.0853^{L}	0.0849^{L}	0.0636	0.0544	0.0108^{C}
skew = 1.75, kurt = 3.75									
1.0	0.0458	0.0352	0.0234^{C}	0.0036^{C}	0.0253	0.0223^{C}	0.0361	0.0170^{C}	0.0030^{C}
.75	0.0740	0.0589	0.0461	0.0166^{C}	0.0651	0.0635	0.0609	0.0375	0.0087^{C}
.57	0.1135^{L}	0.0920^{L}	0.0804^{L}	0.0529	0.1300^{L}	0.1534^{L}	0.1141^{L}	0.0688	0.0226^{C}
.48	0.1153^{L}	0.0863^{L}	0.0763^{L}	0.0503	0.1060^{L}	0.1174^{L}	0.0895^{L}	0.0662	0.0153^{C}
skew = 3.00, kurt = 21.00									
1.0	0.0329	0.0226^{C}	0.0147^{C}	0.0020^{C}	0.0193^{C}	0.0163^{C}	0.0297	0.0096^{C}	0.0018^{C}
.75	0.0643	0.0468	0.0364	0.0137^{C}	0.0590	0.0659	0.0532	0.0262	0.0043^{C}
.57	0.1318^{L}	0.1007^{L}	0.0883^{L}	0.0504	0.1565^{L}	0.2021^{L}	0.1196^{L}	0.0721	0.0177^{C}
.48	0.1144^{L}	0.0757^{L}	0.0667	0.0439	0.1113^{L}	0.1419^{L}	0.0849^{L}	0.0528	0.0126^{C}
n = 15									
skew = 0, kurt = 0									
1.0	0.0496	0.0473	0.0382	0.0048^{C}	0.0455	0.0358	0.0463	0.0364	0.0224^{C}
.75	0.0695	0.0578	0.0482	0.0166^{C}	0.0515	0.0472	0.0516	0.0466	0.0250
.57	0.0809^{L}	0.0551	0.0491	0.0245^{C}	0.0567	0.0570	0.0519	0.0472	0.0219^{C}
.48	0.0827^{L}	0.0490	0.0450	0.0299	0.0519	0.0530	0.0445	0.0433	0.0229^{C}
skew = 1.00, kurt = 0.75									
1.0	0.0492	0.0443	0.0354	0.0054^{C}	0.0397	0.0341	0.0511	0.0338	0.0163^{C}
.75	0.0692	0.0564	0.0489	0.0189^{C}	0.0572	0.0574	0.0590	0.0467	0.0214^{C}
.57	0.0904^{L}	0.0651	0.0592	0.0345	0.0791^{L}	0.0927^{L}	0.0736	0.0532	0.0266
.48	0.0996^{L}	0.0622	0.0579	0.0378	0.0715	0.0812^{L}	0.0649	0.0524	0.0235^{C}
skew = 1.75, kurt = 3.75									
1.0	0.0432	0.0351	0.0269	0.0053^{C}	0.0321	0.0299	0.0458	0.0230^{C}	0.0103^{C}
.75	0.0701	0.0551	0.0452	0.0180^{C}	0.0678	0.0837^{L}	0.0650	0.0410	0.0195^{C}
.57	0.1049^{L}	0.0831^{L}	0.0763^{L}	0.0489	0.1237^{L}	0.1841^{L}	0.1095^{L}	0.0653	0.0371
.48	0.1113^{L}	0.0789^{L}	0.0735	0.0499	0.1016^{L}	0.1442^{L}	0.0828^{L}	0.0650	0.0310
skew = 3.00, kurt = 21.00									
1.0	0.0341	0.0231^{C}	0.0180^{C}	0.0028^{C}	0.0230^{C}	0.0198^{C}	0.0319	0.0114^{C}	0.0040^{C}
.75	0.0668	0.0482	0.0411	0.0165^{C}	0.0679	0.1033^{L}	0.0596	0.0310	0.0127^{C}
.57	0.1174^{L}	0.0937^{L}	0.0865^{L}	0.0547	0.1427^{L}	0.2458^{L}	0.1204^{L}	0.0760^{L}	0.0358
.48	0.1120^{L}	0.0768^{L}	0.0724	0.0494	0.1108^{L}	0.1794^{L}	0.0817^{L}	0.0621	0.0204^{C}
n = 30									
skew = 0, kurt = 0									
1.0	0.0481	0.0472	0.0438	0.0080^{C}	0.0434	0.0402	0.0473	0.0430	0.0380
.75	0.0613	0.0486	0.0440	0.0167^{C}	0.0495	0.0486	0.0507	0.0450	0.0372
.57	0.0764^{L}	0.0512	0.0487	0.0270	0.0498	0.0506	0.0485	0.0491	0.0367
.48	0.0826^{L}	0.0502	0.0492	0.0343	0.0521	0.0533	0.0522	0.0484	0.0433
skew = 1.00, kurt = 0.75									
1.0	0.0525	0.0482	0.0435	0.0070^{C}	0.0433	0.0430	0.0522	0.0428	0.0367
.75	0.0661	0.0528	0.0479	0.0177^{C}	0.0649	0.0681	0.0559	0.0473	0.0390
.57	0.0874^{L}	0.0607	0.0579	0.0362	0.0942^{L}	0.1163^{L}	0.0648	0.0545	0.0416
.48	0.0899^{L}	0.0536	0.0511	0.0347	0.0823^{L}	0.0973^{L}	0.0552	0.0497	0.0400

Table 2 Continued

ε	*F*	HF	Box	GG	.1-Trimmed	.2-Trimmed	T^2	Bootstrap-*F*	Bootstrap-T^2
n = 30									
skew = 1.75, kurt = 3.75									
1.0	0.0466	0.0424	0.0378	0.0072^{C}	0.0371	0.0365	0.0462	0.0362	0.0271
.75	0.0664	0.0533	0.0495	0.0190^{C}	0.0844^{L}	0.1109^{L}	0.0576	0.0470	0.0379
.57	0.0944^{L}	0.0671	0.0637	0.0385	0.1662^{L}	0.2501	0.0809^{L}	0.0560	0.0456
.48	0.1042^{L}	0.0672	0.0645	0.0449	0.1280^{L}	0.1836^{L}	0.0673	0.0592	0.0408
skew = 3.00, kurt = 21.00									
1.0	0.0425	0.0306	0.0261	0.0046^{C}	0.0294	0.0267	0.0413	0.0198^{C}	0.0123^{C}
.75	0.0686	0.0483	0.0452	0.0157^{C}	0.1083^{L}	0.1657^{L}	0.0603	0.0389	0.0225^{C}
.57	0.1095^{L}	0.0811^{L}	0.0783^{L}	0.0483	0.2106^{L}	0.3457^{L}	0.1052^{L}	0.0710	0.0493
.48	0.1056^{L}	0.0696	0.0673	0.0480	0.1602^{L}	0.2553^{L}	0.0826^{L}	0.0611	0.0315
n = 60									
skew = 0, kurt = 0									
1.0	0.0538	0.0528	0.0511	0.0109^{C}	0.0472	0.0460	0.0542	0.0516	0.0463
.75	0.0637	0.0478	0.0461	0.0187^{C}	0.0457	0.0450	0.0558	0.0479	0.0490
.57	0.0844^{L}	0.0538	0.0521	0.0300	0.0477	0.0464	0.0514	0.0530	0.0419
.48	0.0924^{L}	0.0537	0.0525	0.0370	0.0525	0.0544	0.0535	0.0532	0.0483
skew = 1.00, kurt = 0.75									
1.0	0.0478	0.0462	0.0441	0.0078^{C}	0.0471	0.0436	0.0494	0.0443	0.0483
.75	0.0672	0.0521	0.0496	0.0204^{C}	0.0713	0.0830^{L}	0.0525	0.0496	0.0462
.57	0.0789^{L}	0.0523	0.0509	0.0282	0.1156^{L}	0.1544^{L}	0.0545	0.0497	0.0471
.48	0.0925^{L}	0.0562	0.0548	0.0382	0.0878^{L}	0.1104^{L}	0.0562	0.0540	0.0493
skew = 1.75, kurt = 3.75									
1.0	0.0490	0.0454	0.0426	0.0071^{C}	0.0414	0.0397	0.0497	0.0431	0.0397
.75	0.0692	0.0550	0.0528	0.0185^{C}	0.1059^{L}	0.1580^{L}	0.0579	0.0519	0.0434
.57	0.0861^{L}	0.0587	0.0565	0.0358	0.2217^{L}	0.3648^{L}	0.0687	0.0528	0.0485
.48	0.0937^{L}	0.0555	0.0550	0.0397	0.1640^{L}	0.2563^{L}	0.0610	0.0510	0.0465
skew = 3.00, kurt = 21.00									
1.0	0.0470	0.0393	0.0368	0.0081^{C}	0.0352	0.0355	0.0473	0.0330	0.0207^{C}
.75	0.0732	0.0549	0.0529	0.0207^{C}	0.1348^{L}	0.2411^{L}	0.0592	0.0484	0.0302
.57	0.0940^{L}	0.0654	0.0636	0.0409	0.2903^{L}	0.5217^{L}	0.0836^{L}	0.0591	0.0508
.48	0.0942^{L}	0.0580	0.0574	0.0405	0.2172^{L}	0.3922^{L}	0.0697	0.0542	0.0399

Note. HF = Huynh-Feldt adjusted *F* test; Box = Box adjusted *F* test; GG = Geisser-Greenhouse adjusted *F* test. C. Conservative performance, denoting values that fall below the interval .025-.075 (Bradley's criterion). L. Liberal performance, denoting values that fall above the interval .025-.075 (Bradley's criterion).

approach in a few specific cells out of 64, the relative performance of the methods overall was largely the same as that using Bradley's liberal criterion; for this reason, the additional results for the confidence intervals are not reported here.

Preliminary general observations are as follows, created by averaging information from Table 2. When averaged across sample sizes and distributional conditions, Type I error rates tend to become increasingly liberal as nonsphericity increases, although it is interesting that all methods but the unadjusted *F* and the GG *df*-adjusted *F* actually decrease slightly in Type I error rate (on average) from ε = .57 to ε = .48, which may possibly be due to the specific correlational structure used to create these levels of nonspheric-

Table 3
Frequencies of Cells (out of 64) Whose Type I Error Rate Falls Outside Bradley's Robustness Interval

Method	Conservative	Liberal	Total	Conservative	Liberal	Total
	Normal/Spherical (4 cells)			Nonnormal/Spherical (12 cells)		
F	0	0	0	0	0	0
HF	0	0	0	2	0	2
Box	0	0	0	3	0	3
GG	4	0	4	12	0	12
.1-Trimmed	0	0	0	2	0	2
.2-Trimmed	0	0	0	3	0	3
T^2	0	0	0	0	0	0
Bootstrap *F*	0	0	0	5	0	5
Bootstrap T^2	2	0	2	8	0	8
	Normal/Nonspherical (12 cells)			Nonnormal/Nonspherical (36 cells)		
F	0	8	8	0	24	24
HF	0	0	0	0	9	9
Box	0	0	0	0	6	6
GG	6	0	6	12	0	12
.1-Trimmed	0	0	0	0	27	27
.2-Trimmed	0	0	0	0	30	30
T^2	0	0	0	0	13	13
Bootstrap *F*	0	0	0	0	1	1
Bootstrap T^2	5	0	5	15	0	15

Note. Smaller values indicate better performance.

ity. Similarly, when averaged across sample sizes and sphericity conditions, increased nonnormality tends to yield increased Type I error rates, although from moderate to severe nonnormality some methods stabilize or decrease slightly. Finally, when averaged across sphericity and normality conditions, behavior seems to differ across methods. In general, all the methods except the unadjusted *F*, GG, and both levels of the β-trimmed means performed well with large sample sizes of $n = 60$. Each method will now be considered individually.

Univariate Methods

The *traditional unadjusted F* behaves properly when conditions are normal and spherical. When nonnormality is introduced but sphericity is preserved, the *F* test still remains within the robustness interval. When normality is preserved but nonsphericity is introduced, the Type I error rates become systematically more liberal until in the two most extreme nonspherical cases

the unadjusted *F* becomes liberal under all sample sizes. Finally, under violations of both normality and sphericity, the results are identical to those for sphericity alone: Only those cases with $\varepsilon = .57$ and $\varepsilon = .48$ show unacceptably liberal control over Type I error.

As for the *GG* df-*adjustment*, under normal and spherical conditions it is, not surprisingly, conservative in its control at all sample sizes. It remains similarly conservative with the introduction of nonnormality when spherical conditions are preserved. When nonsphericity is present but distributions are normal, the performance of the GG method varies with ε and *n*. Specifically, it is still conservative with $\varepsilon = .75$ under all sample sizes. In the more extreme condition of $\varepsilon = .57$ the GG method returns to the confines of the robustness interval given sufficient sample size ($n = 30$ and $n = 60$), whereas the smaller sample size conditions are not able to overcome the conservatism. In the most nonspherical condition, GG remains robust at any sample size. This may seem counterintuitive, but remember that this method is a lower-bound correction assuming the worst possible sphericity condition—1/(4 – 1) or .3333 in this case. In all sphericity conditions investigated here, this correction is technically too strong, but as sphericity approaches this lower bound level, the error rate tends to lose its conservatism and rises into the robustness interval. When nonnormality is coupled with nonsphericity, results are conservative under all distributions for all sample sizes when $\varepsilon = .75$; for more extreme sphericity results are within the robustness interval under all distributions for all sample sizes.

Considering *Box's* $\hat{\varepsilon}$ df-*adjustment*, it is neither conservative nor liberal under normal and spherical conditions. When nonnormality is present for spherical data, it becomes conservative only for the most extreme nonnormality when $n = 15$ and for the two most extreme nonnormal conditions when $n = 10$. On the other hand, when data are normal but sphericity is violated, Box's method was robust at all sphericity levels and sample sizes. As for the combination of nonnormality and nonsphericity, this method is robust under $n = 60$. With $n = 30$, the most nonnormal scenario produces a liberal result with $\varepsilon = .57$, although it is interesting that the error rate drops back into the robustness region as ε reaches the most extreme level of .48. For sample sizes of $n = 15$, liberal results are observed under the two most nonnormal conditions with $\varepsilon = .57$; again, the method is robust at the most extreme $\varepsilon = .48$. Finally, for $n = 10$ the two most extremely nonnormal conditions yield a liberal result with $\varepsilon = .57$; curiously, only the less nonnormal of the two yields liberal results in the $\varepsilon = .48$ case. It would appear that the relationship of distributional form and sphericity with Type I error rate is interactive and occasionally nonmonotonic in nature.

The *HF* $\tilde{\varepsilon}$ df-*adjustment* is also robust when normality and sphericity are present. When data become nonnormal but spherical, HF remains robust except under the most extreme nonnormal situations when *n* is as small as 15

and 10, in which cases it becomes conservative. With normal data under nonspherical conditions, HF remains completely robust under all sample sizes and all sphericity conditions. When nonnormality and nonsphericity are combined, results become considerably more complex. For the largest sample size, all results fall within the robustness interval. When $n = 30$, only the most extreme nonnormality exhibits nonrobustness; specifically, $\varepsilon = .57$ yields liberal control, whereas the most extreme $\varepsilon = .48$ remains robust. For both $n = 15$ and $n = 10$, the weakest nonnormal condition shows robustness at all sphericity levels; with more extreme nonnormality, error rates become liberal for all cases where $\varepsilon = .57$ and .48.

The results for the β-*trimmed means* are similar for the two levels, with the .1-trimmed mean performing slightly better than the .2-trimmed mean. Although neither is conservative nor liberal when the data are normal under spherical and nonspherical conditions, when the data are nonnormal and spherical the .1-trimmed mean statistic performs conservatively in the most extreme nonnormal condition when $n = 10$ and $n = 15$. The .2-trimmed mean performs conservatively under these conditions, and in addition, in the less extreme nonnormal condition when $n = 10$. In the combined nonnormal and nonspherical conditions, results are liberal for both levels of the β-trimmed mean in the two most extreme nonspherical conditions across all levels of nonnormality for $n = 10$. As sample size increases, the Type I error rates also increase for the tests of these statistics. At $n = 60$, with the exception of the .1-trimmed mean at $\varepsilon = .75$ at the lowest level of nonnormality, the trimmed-mean statistics are liberal at all levels of nonnormality combined with nonsphericity. The most extreme example is the .2-trimmed mean statistic, which reaches a Type I error rate of more than 50% in the most nonnormal condition with $\varepsilon = .57$.

The last of the methods based on the univariate F is the *bootstrap-F* approach, which remains within the robustness interval at all sample sizes when normality and sphericity are present. Error rates actually appear to decline as n gets smaller, but they never fall below the defined limits of robustness. With the introduction of nonnormality under spherical conditions, results appear robust for the largest sample size of $n = 60$. When $n = 30$, results for the most extreme nonnormality become conservative. For both of the smallest sample sizes, error rates are conservative in the two most nonnormal scenarios. If normality is preserved while nonsphericity is increased, the bootstrap-F remains robust under all levels of ε and with all sample sizes. Finally, with nonnormal and nonspherical data, the bootstrap-F approach is robust under all conditions except for when it becomes liberal in the most extreme distribution with $n = 15$ and $\varepsilon = .57$ (but not with a smaller sample size or with worse sphericity). Given the previous result that nonnormality alone can yield conservative results, the fact that additional

nonsphericity seems to temper that nonrobustness may be indicative of a complex interaction taking place between distributional form and sphericity.

Multivariate Methods

The *traditional multivariate* T^2 test appears remarkably robust when data are normal and/or spherical; that is, the violation of none or one of these conditions does not yield a result outside the robustness interval for any sample size. However, when nonnormality and nonsphericity are present, this method quickly becomes one of the worst performing options investigated. With the largest sample size $n = 60$, the most extreme nonnormality shows a liberal result when $\varepsilon = .57$. When $n = 30$, liberal results are observed when $\varepsilon = .57$ in the two most extremely nonnormal distributions, and when $\varepsilon = .48$ for the most extremely nonnormal distribution. With sample size equal to 15, the two most nonnormal distributions show liberal results for the two most nonspherical conditions. And for the smallest sample size of $n = 10$, the same pattern of results is observed as with $n = 15$ plus an additional liberal outcome when the case of $\varepsilon = .57$ under the mildest nonnormal distribution.

The *bootstrap-*T^2 method performs conservatively in its Type I error control when sample sizes are small. For $n = 10$, the bootstrap-T^2 method is extremely conservative under *all* conditions. In addition, between .1% and 1% of the bootstrapped samples need to be replaced due to near singular covariance matrices for this small sample size. As sample size increases to $n = 15$, fewer than .01% of the bootstrapped samples needed replacement for any condition. The bootstrap-T^2 method performs more conservatively as the data become more spherical, and it becomes more conservative as nonnormality becomes more severe. These opposite effects lead to an interesting result when $n = 15$ with the bootstrap-T^2 appearing robust under combined moderate to severe violations of both normality and sphericity, but performing conservatively in all the cases where the data are spherical ($\varepsilon = 1.0$), as well as performing conservatively in the condition with the most severe violation of both assumptions. When $n = 30$ or $n = 60$, none of the bootstrapped data sets needed to be replaced, and the bootstrap-T^2 method mostly performs within Bradley's robustness criterion. With these larger sample sizes, the bootstrap-T^2 is conservative for the most severely nonnormal condition when $\varepsilon = 1.0$ and $\varepsilon = .75$ for $n = 30$, and $\varepsilon = 1.0$ for $n = 60$. It was hoped that the bootstrap-T^2 would retain the positive attributes of T^2 under nonnormal or nonspherical conditions while offering greater robustness when both threats are present. Such was not the case, as results show that the bootstrap-T^2 method performs conservatively under most conditions.

Conclusions

For a one-way repeated measure design with a levels, clearly no one procedure for testing H_0: $\mu_1 = ... = \mu_a$ works well under all situations. The current investigation, which included many combinations of sample size, distributional properties, and sphericity conditions, leads to the conclusion that a repeated measure ANOVA should be preceded by more than just a test for sphericity such as Mauchly's test. Preparatory work should also include a test for nonnormality (e.g., a Kolmogorov-Smirnov test). If neither violation appears to be present, all but the GG and the bootstrap-T^2 method seem to be reasonable options. Perhaps the best alternatives are the traditional unadjusted F or the one-sample multivariate T^2, which take full advantage of the assumptions of normality and sphericity. If a violation of sphericity occurs alone, the HF, Box, T^2, β-trimmed mean method, and bootstrap-F all appear to offer reasonable Type I error control. If nonnormality occurs as the sole violation, both the traditional F and multivariate T^2 appear most robust (even though they are technically dependent on the assumption of normality). If violations of both normality and sphericity occur, the bootstrap-F method seems to be far and away the most robust alternative, even with fairly small sample sizes.

The above recommendations can be simplified further. If data are normal and/or spherical, one may simply use a one-sample multivariate T^2 test. If data are neither normal nor spherical, the bootstrap-F method proposed herein should be used. These suggestions are based solely on Type I error control; further investigations examining the statistical power of methods appearing to control Type I error satisfactorily could possibly temper these recommendations. Notwithstanding, the bootstrap-F method appears quite promising for testing data under the twin threats of nonnormality and nonsphericity. Further research would be useful to extend the bootstrapping approach to multifactor repeated measure designs as well as to split-plot designs.

Appendix
Summary Information for Simulated 100,000 × 4 Verification Data Matrix

Coefficient	Target	Estimate	Target	Estimate	Target	Estimate	Target	Estimate
skew	0.0000	–0.0071	0.0000	–0.0035	0.0000	0.0044	0.0000	0.0022
kurtosis	0.0000	0.0016	0.0000	0.0054	0.0000	–0.0098	0.0000	–0.0043
ε	1.0000	1.0000	0.7500	0.7483	0.5700	0.5662	0.4800	0.4791
skew	1.0000	1.0017	1.0000	1.0073	1.0000	1.0076	1.0000	1.0089
kurtosis	0.7500	0.7578	0.7500	0.7822	0.7500	0.7831	0.7500	0.7698
ε	1.0000	1.0000	0.7500	0.7497	0.5700	0.5665	0.4800	0.4777
skew	1.7500	1.7651	1.7500	1.7302	1.7500	1.7484	1.7500	1.7598
kurtosis	3.7500	3.8618	3.7500	3.5895	3.7500	3.7120	3.7500	3.8002
ε	1.0000	1.0000	0.7500	0.7448	0.5700	0.5690	0.4800	0.4797
skew	3.0000	3.1218	3.0000	3.0677	3.0000	3.0200	3.0000	2.8957
kurtosis	21.0000	22.5303	21.0000	22.7421	21.0000	20.3486	21.0000	20.0559
ε	1.0000	0.9999	0.7500	0.7543	0.5700	0.5688	0.4800	0.4774

Note. Skew and kurtosis estimates are averaged across simulated data from all four levels of the repeated measure (i.e., 400,000 pieces of data in each case), whereas the sphericity estimate is from the 4 × 4 covariance matrix based on 100,000 cases.

References

Aptech Systems. (1996). *GAUSS system and graphics manual*. Maple Valley, WA: Aptech Systems, Inc.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems. *Annals of Mathematical Statistics*, *25*, 290-302, 484-498.

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

DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological Methods*, *2*, 292-307.

Diaconis, P., & Efron, B. (1983, May). Computer-intensive methods in statistics. *Scientific American*, 116-130.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, *37*, 36-48.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

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

Harwell, M. R., & Serlin, R. C. (1997). An empirical study of five multivariate tests for the single factor repeated measures model. *Communications in Statistics-Simulation and Computation*, *26*, 605-618.

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

Keppel, G. (1991). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice Hall.

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

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1994, April). *Analysis of single-group repeated measures designs: A quantitative review.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Lomax, R. G. (1998). *Statistical concepts: A second course for education and the behavioral sciences*. Mahwah, NJ: Lawrence Erlbaum.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. Pacific Grove, CA: Duxbury.

Lunneborg, C. E., & Tousignant, J. P. (1985). Efron's bootstrap with application to the repeated measures design. *Multivariate Behavioral Research*, *20*, 161-178.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data*. Belmont, CA: Wadsworth.

Nevitt, J., & Hancock, G. R. (1999). PWRCOEFF & NNORMULT: A set of programs for simulating multivariate nonnormal data. *Applied Psychological Measurement*, *23*, 54.

Quintana, S. M., & Maxwell, S. E. (1994). A Monte Carlo comparison of seven ε-adjustment procedures in repeated measures designs with small sample sizes. *Journal of Educational Statistics*, *19*, 57-71.

Rasmussen, J. L. (1987). Parametric and bootstrap approaches to repeated measures designs. *Research Methods, Instruments and Computers*, *4*, 357-360.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum.

Tandon, P. K., & Moeschberger, M. L. (1989). Comparison of nonparametric and parametric methods in repeated measures designs—A simulation study. *Communications in Statistics-Simulation and Computation*, *18*, 777-792.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate non-normal distributions. *Psychometrika*, *48*, 465-471.

Westfall, P. H., & Young, S. S. (1993). *Resampling based multiple testing*. New York: John Wiley.

Wilcox, R. R. (1993). Analyzing repeated measures or randomized block designs using trimmed means. *British Journal of Mathematical and Statistical Psychology*, *46*, 63-76.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*, 300-314.