Journal of Modern Applied Statistical Methods

Volume 12 | Issue 2 Article 4

11-1-2013

A Monte Carlo Comparison of Robust MANOVA Test Statistics

Holmes Finch
Ball State University, Muncie, IN, whfinch@bsu.edu

Brian French
Washington State University, Pullman, WA, frenchb@wsu.edu

Journal of Modern Applied Statistical Methods
November 2013, Vol. 12, No. 2, 35-81.

ISSN 1538 − 9472

Regular Articles: A Monte Carlo Comparison of Robust MANOVA Test Statistics

Holmes Finch
Ball State University
Muncie, IN

Brian French
Washington State University
Pullman, WA

Multivariate Analysis of Variance (MANOVA) is a popular statistical tool in the social sciences, allowing for the comparison of mean vectors across groups. MANOVA rests on three primary assumptions regarding the population: (a) multivariate normality, (b) equality of group population covariance matrices and (c) independence of errors. When these assumptions are violated, MANOVA does not perform well with respect to Type I error and power. There are several alternative test statistics that can be considered including robust statistics and the use of the structural equation modeling (SEM) framework. This simulation study focused on comparing the performance of the *P* test statistics with fifteen other test statistics across seven manipulated factors. These statistics were evaluated across 12,076 different conditions in terms of Type I error and power. Results suggest that when assumptions were met, the standard MANOVA test functioned well. However, when assumptions were violated, it performed poorly, whereas several of the alternatives performed better. Discussion focuses on advice for selecting alternatives in practice. This study's focus on all these in one simulation and the 3 group case should be helpful to the practitioner making methodological sections.

Keywords: MANOVA, robust statistics, structural equation modeling, nonparametric, mean comparisons, Monte Carlo simulation

Introduction

Much research in the social sciences involves the comparison of means for two or more groups across multiple related outcome measures. For example, studies examining the impact of interventions on multiple measures of academic, social, communication, and emotional development are common in education and psychology. Parenting our Children to Excellence (*PACE*) (Dumas et al., 1999) is

Dr. Finch is a Professor of psychometrics and statistics in the Department of Educational Psychology. Email him at: whfinch@bsu.edu. Dr. French is a professor of measurement and psychometrics in the Department of Educational Leadership and Counseling Psychology. Email him at: frenchb@wsu.edu.

such an intervention project that has been tested through randomized control trials evaluating an 8-week program that teaches positive parenting techniques aimed at increasing parenting skills and child positive behavior. In programs such as this, there are typically multiple correlated outcome variables (e.g., child disruptive behaviors, child adjustment, parenting behaviors, parenting competence), which can have high-stakes implications (e.g., resource allocation, curriculum development, policy decisions). Therefore, given that high stakes decisions may be based upon the results of statistical analyses, precise modeling of data is paramount.

This type of research design in intervention work may revolve around hypotheses regarding group differences on a set of variables, rather than on individual variables. Multivariate hypotheses lead a researcher to a multivariate analysis, as it may be most appropriate for assessing group differences on the set of variables (Huberty & Olejnik, 2006). Specifically, multivariate analysis of variance (MANOVA) is well-suited for testing hypotheses about differences between groups (Hair, Anderson, Tatham, & Black, 1987). MANOVA can be viewed as a direct extension of the univariate general linear model that is most appropriate for examining differences between groups on several variables simultaneously (Hair et al., 1987; Olejnik, 2010). As Hancock, Lawrence and Nevitt (2001) pointed out, “MANOVA evaluates group differences on a linear composite of observed variables constructed so as to maximally differentiate the groups in multivariate space" (p. 535).

Situations are described here in which MANOVA may be the optimal analysis (particularly when compared with univariate analysis of variance (ANOVA)). Following this discussion, particular data structures that may cause problems for MANOVA will be described, particularly when key assumptions are violated, and then several approaches for dealing with the assumption violations. A simulation study comparing these methods across a variety of conditions is reported, and conclude the discussion with recommendations for researchers using MANOVA in cases where the assumptions are not met.

Despite the fact that MANOVA may be the optimal analysis for a multivariate problem due to its relative ease of use and interpretation, researchers may often employ multiple independent ANOVA models to determine if there are significant differences among group means on each of several outcome measures of interest. In the previous example with PACE, five separate ANOVAs could be conducted to determine if the treatment and control groups differed on the related outcomes. Although this approach may be familiar to many researchers, the simplicity of the univariate ANOVA could also lead to unwarranted conclusions

due to inflation of the family-wise Type I error rate and a potential decrease in power when the response is actually multivariate in nature. In fact, McCarroll, Crays, & Dunlap (1992) provided evidence that Type I error rates are inflated when ANOVA is used in a sequential manner. For example, the family-wise Type I error rate for testing the 5 outcomes in the PACE data, assuming alpha = 0.05, would be 0.23. It is acknowledged that by adjusting critical values for the univariate situation, the Type I error rate can be controlled (Ramsey, 1982). In fact, Ramsey illustrated that the Bonferroni procedures showed greater robustness in many cases compared to methods based on Hotelling's T^2 statistic, which requires more and stronger assumptions (e.g., multivariate normality) compared to Bonferroni procedures.

Often the research question of interest concerns differences on a set of related or correlated outcome variables, not each variable separately. That is, the researcher wants to examine questions about how groups differ along a combination of correlated dimensions or variables, not one dimension or variable at a time. Univariate procedures cannot provide insight on the former, as each variable is examined in isolation. As a result of this inability to consider the entire multivariate response space, the practice of following up a significant MANOVA result with individual ANOVAs does not provide insight to questions regarding multivariate differences (e.g., Huberty & Morris, 1989). Harris (2001) suggested that the use of MANOVA for between-group comparisons is more appropriate in the context of multiple dependent variables compared to the use of many individual univariate tests.

There is recognition that MANOVA may not be the best choice in all cases in which multiple outcome variables are of interest. The choice of the analytic procedure does rest on several factors including the data, research design, and research questions. For example, if the outcome variables are uncorrelated or have high positive correlations, then MANOVA may not be as effective as conducting separate univariate ANOVAs (Tabachnick & Fidell, 2007). In contrast, MANOVA can have greater power compared to the univariate methods when there is a moderate to strong negative correlation between the dependent variables (Tabachnick & Fidell, 2007). Additionally, power can depend on the relationship between dependent variables and the effect size (Cole, Maxwell, Arvey, & Salas, 1994). This study focuses on situations for which MANOVA may be most appropriate, based on recommendations from the works cited above, and considers the intercorrelations and effect sizes and how they relate to power of several test statistics as well as violations of assumptions, in order to highlight the

performance of these various test statistics associated with MANOVA, under different conditions.

To summarize the discussion heretofore, the decision regarding whether to select a univariate or multivariate comparison of between groups means must be made based on both statistical and substantive considerations. If the research questions are essentially multivariate in nature (e.g. Do the groups differ on the set of dependent variables?) then MANOVA is preferred to ANOVA (Stevens, 2001). In addition, when the dependent variables are at least moderately correlated, MANOVA will generally yield greater power compared to the univariate alternatives. Conversely, if the research questions are focused on the individual variables (e.g. Do the groups differ on Y1? Do the groups differ on Y2?), and/or if the dependent variables have little or no correlation or very strong positive correlations among them, then use of individual ANOVAs rather than MANOVA may be most appropriate (Stevens, 2001). In conclusion, the advantages of MANOVA, beyond Type I error control, can include (a) improving power for identifying group differences, (b) observing differences possibly missed in single ANOVAs (Huberty & Morris, 1989; Tabachnick & Fidell, 2007), and (c) understanding the outcome variables as a system rather than isolated measurements (Huberty & Morris, 1989). This study was conducted to examine performance of the several MANOVA test statistics in the case where multivariate questions are of primary interest and the multivariate procedure would be preferred.

Standard parametric multivariate means comparisons

In evaluating multivariate mean differences with MANOVA in the 2 group case, researchers test the null hypothesis of no group mean vector differences using Hotelling's T^2 statistic. Please see Johnson & Wichern (2002) for additional information on these multivariate test statistics. Hotelling's T^2 statistic which takes the form:

$$T^2 = \left(\bar{Y}_1 - \bar{Y}_2\right)' \left[S\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right]^{-1} \left(\bar{Y}_1 - \bar{Y}_2\right) \quad (1)$$

Where

$\bar{\boldsymbol{Y}}_1 =$ Mean vector for group 1

$\bar{\boldsymbol{Y}}_2 =$ Mean vector for group 2

$n_1 =$ Sample size for group 1

$n_2 =$ Sample size for group 2

$\boldsymbol{S} =$ Sample pooled covariance matrix; $\dfrac{(n_1-1)S_1+(n_2-1)S_2}{n_1+n_2-2}$

$S_1 =$ Covariance matrix for group 1

$S_2 =$ Covariance matrix fro group 2

In this equation, the transpose (') operator is used to create sums of squared differences, in the context of matrices, and the inverse (-1) is used for matrix division. Hotelling's T^2 has been extended to accommodate the case of more than two groups with four different F approximation tests: Pillai's trace, (P) Wilk's lambda (Λ), Hotelling-Lawley Trace (H) and Roy's Greatest Root (R). These test statistics can be expressed as follows:

$$\Lambda = \frac{\lfloor W \rfloor}{\lfloor W+B \rfloor}$$

where (2)

$W =$ within group sum of squares and cross products matrix

$B =$ between group sum of squares and cross products matrix

$$P = tr\left[B(B+W)^{-1} \right] \quad (3)$$

$$H = tr\left[BW^{-1} \right] \quad (4)$$

$R =$ maximum eigenvalue of $W(B+W)^{-1}$ (5)

$|W| =$ Determinant of matrix W, where the determinant can be viewed as generalized or total variance of that matrix

Prior research regarding standard MANOVA test statistic performance

Accurate use and interpretation of these multivariate test statistics is dependent upon the assumptions of independent errors, multivariate normality, and homogeneity of group covariance matrices. When these assumptions are met, the tests perform similarly well with respect to controlling Type I error rates and maintaining appropriate statistical power, particularly in studies with relatively large sample sizes (e.g., Blair, Higgins, Karniski & Kromrey, 1994; Hopkins & Clay, 1963; Johnson & Wichern, 2002; Ramsey, 1982; Stevens, 2001). Several works cited in this review have informed multivariate researchers on how these statistics perform under various conditions. However, this work has primarily

been focused on the 2 group case. In addition, some of this work, particularly Ramsey, treated the data in a univariate fashion, rather than testing multivariate hypotheses about group means on several dependent variables simultaneously. Though this may be appropriate in some cases, many times where multivariate data are present, the hypothesis of interest concerns group differences on the set of means rather than on the individual means, in which case such univariate treatment of the data may be inappropriate (Huberty & Olejnik, 2006).

The work presented here focuses on the situation where researchers are interested in conducting multivariate means testing (rather than univariate), and is unique as (a) many test statistics are compared in a single simulation study, including a latent variable approach, and (b) the 3 group case is considered to ascertain whether the results from the 2 group case can generalize to the 3 group case, certainly a more complex but also perhaps more realistic condition. Many of these methods have been examined in simulation studies. However, the methods included here have not all been examined in a single study. Therefore, though it has been possible to describe how two or three of these statistics perform relative to one another, this study allows for the comparison of all of these methods under the same conditions.

Violations in assumptions of multivariate normality and homogeneity of covariance are often characteristic of social science research, and standard parametric MANOVA has limitations under such conditions (Blair et al., 1994; Everitt, 1979; Finch, 2005). Investigations of Type I error rates and power have suggested that these multivariate tests may not perform well when there are violations in assumptions of multivariate normality and equality of covariance matrices (e.g., Hakstian, Roed & Lind, 1979; Hopkins & Clay, 1963; Olson, 1974; Lee, 1971; Pillai & Jayachandran, 1967). Perhaps most notable is the performance of Hotelling T^2 in studies of unequal sample sizes when the assumptions of multivariate normality and particularly equality of covariance matrices has not been met. In such cases, the T^2 demonstrated diminished power as the degree of skewness of the response variables increased (Everitt, 1979). Furthermore, when the groups' covariance matrices were not homogeneous, the Type I error rate of the T^2 was inflated when the groups were not of equal size and the smaller group had the larger variances (Hakstian, Roed & Lind, 1979; Hopkins & Clay, 1963).

These results for T^2 are similar to those reported in studies of the performance of Pillai's Trace, Wilk's Lambda, Hotelling-Lawley's Trace and Roy's Greatest Root when there are violations in the assumption of equality of covariance matrices (Finch, 2005; Olson, 1974; Sheehan-Holt, 1998). In these

studies, when the smaller group had the larger variance the Type I error rates were inflated, whereas when the larger group had the larger elemental covariance elements, there was a reduction in power. Non-normality characterized by relatively severe skewness also resulted in a reduction of power (Everitt, 1979; Finch, 2005). Furthermore, when the assumptions were violated, Pillai's Trace was relatively more robust in terms of Type I error rate control compared to Wilk's Lambda and Hotelling-Lawley's Trace but exhibited somewhat lower power compared to these other tests. Not one of the common MANOVA statistics can be clearly identified as the single best test for use in all situations (Lee, 1971; Pillai & Jayachandran, 1967). The comparative effectiveness of these methods changed relative to specific features of the data. However, taken across a broad sweep of real data conditions, Λ, P and H all generally perform similarly, particularly when standard assumptions are met (Johnson & Wichern, 2002). In summary, the standard test statistics used with MANOVA are deleteriously affected by violations of the assumptions of normality and homogeneity of covariance matrices, particularly when samples are of unequal sizes.

Alternative test statistics to standard MANOVA when assumptions are violated

In response to these problems associated with assumption violations, a number of alternative test statistics have been investigated, particularly for use in the absence of multivariate normality and when group covariance matrices are not equal. The formulas for many of the basic versions of these statistics appear in Appendix A for the interested reader. Table 1 provides summary information across the different statistical tests to assist with organizing the information.

Brown and Forsythe (1974), James (1954), Johansen (1980), Yao (1965) and Nel and van der Merwe (1986) each outlined alternatives to the standard multivariate test statistic in the presence of unequal covariance matrices. Extensions of Hotelling's T^2, these parametric multivariate alternatives examine multiple outcomes between two groups, and have been extended for use with more than two groups. In the two groups case, the James (F_{JA}), Johansen (F_{JN}), Nel and van der Merwe (F_{NV}), and Yao (F_Y) statistics are based on the multivariate analog of the univariate *t*-test equation for unequal variances. $T^2_{unequal} = \left(\bar{Y}_1 - \bar{Y}_2\right)' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} \left(\bar{Y}_1 - \bar{Y}_2\right)$ As with the univariate version of this statistic, the group variances (covariance matrices in the multivariate context) are

Table 1. General Conclusions based on the Literature of Test Statistics Examined for MANOVA Under Various Assumptions Conditions

	Assumptions	
Statistic	***Met***	***Not Met***
Standard (*P, H, L*)	Type I error rate controlled; Optimal power	Inflated Type I error for unequal covariance matrices and reduction of power for severely skewed data
F_{JA}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices; low power for small ratios of sample size to number of dependent variables. Not robust to non-normal data.
F_{JN}, F_{NV}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices; low power for small ratios of sample size to number of dependent variables. Not robust to non-normal data.
F_{Y}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices; low power for small ratios of sample size to number of dependent variables. Not robust to non-normal data.
F_{BF}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices; low power for small ratios of sample size to number of dependent variables. Not robust to non-normal data.
F_{K}	Comparable results to the standard test statistic.	Robust to unequal covariance matrices but not to non-normal data.
TF_{J}, TF_{JN}	Comparable results to the standard test statistic.	For skewed and heavy tailed data, displayed higher power and better Type I error control than did F_{JN}.
TF_{NV}, TF_{Y}, TF_{BF}, TF_{K}	Comparable results to the standard test statistic.	For skewed and heavy tailed data, displayed higher power than did F_{K}.
Rank based test	Comparable Type I error rates to standard test but lower power.	For unequal covariance matrices, displayed better Type I error control though rates were still inflated.
SEM	Comparable Type I error and power rates to standard test for samples of 100 or greater.	Better Type I error control and higher power rates than standard tests for unequal covariance matrices

Note: T^2 = Hotelling's (1931); BF = Brown&Forsythe (1974), J = James (1954); JN = Johansen (1980),K= Kim(1992);NV= Nel & van der Merwe (1986),Y=Yao (1965), SEM = Structural Equation Modeling (Raykov, 2001), T with test = trimmed.

not pooled. The difference between F_{JA} and F_{JN} is in the way that they determine the critical value for assessing statistical significance. The F_{JA} statistics is simply $T^2_{unequal}$ (See Appendix A) with the critical value based on the χ^2 distribution adjusted by a complex term involving the traces of the covariance matrices for the

two groups. In contrast, the value for F_{JN} involves the conversion of $T^2_{unequal}$ to an *F* value, as seen in Appendix A.

The F_{NV} test statistic also is a transformed version of $T^2_{unequal}$ (see Appendix A) and compared to a critical *F* value. Krishnamoorthy and Xia (2006) presented a modified version of the degrees of freedom for F_{NV} labeling their statistic the Modified Nel and van der Merwe test (F_{MNV}). The test statistic remains the same, but the resulting value is compared to a critical *F* value with *p*, v_{KX} degrees of freedom, and the resulting test is affine invariant (results of the test are invariant under a linear transformation of the data). For a more thorough treatment of the calculation of v_{KX} the interested reader is encouraged to read Krishnamoorthy and Xia. Finally, among this set of statistics based upon the $T^2_{unequal}$ value is Yao's F_Y, which incorporates a different weighting scheme involving the determinant of the ratio of group covariance matrices (See Appendix A). Given that these previously described methods share a common root, namely $T^2_{unequal}$, they are discussed as a set of test statistics (i.e., Family 1). An examination of Appendix A reveals that although these statistics share a common root, they vary in terms of how they weight the groups' covariance matrices, and how degrees of freedom are calculated.

Of the alternatives to the standard T^2 described here, the Brown and Forsythe (F_{BF}) and the Kim (F_K) tests are not based on the $T^2_{unequal}$ statistic. The centerpiece of F_{BF} is T_{BF}, which differs from the $T^2_{unequal}$ statistic in terms of how the group covariance matrices are weighted, as can be seen in Appendix A. Essentially, where $T^2_{unequal}$ weights them by the inverse of sample size, T_{BF} uses the proportion of the total sample *not* in a specific group as the weight. Otherwise, T_{BF} is generally similar to $T^2_{unequal}$. The F_{BF} statistic is then compared to the critical value $F_{vBF1,vBF2}$. Kim's (F_K) statistic also is based on an alternative to $T^2_{unequal}$ and is compared with the $F_{m,vk}$ critical value. The calculation for F_K can be found in Appendix A. In general, it differs from both $T^2_{unequal}$ and T_{BF} in the way in which the group covariance matrices are weighted and combined. A review of Appendix A demonstrates that F_K relies on a more complex weighting system to combine these covariance matrices, using as a weight the determinant of their ratio (in the simplest two groups case) raised to the 1/(2*number of predictor variables) power. To simplify further discussion, and given their

similarity in terms of calculation, as mentioned previously F_{JN}, F_{NV}, F_{Y}, and F_{JA} have been organized into one family (Family 1) of statistics, and F_{BF} and F_{K} constitute a another family of test statistics (Family 2).

Prior research regarding alternative MANOVA test statistic performance

The test statistics in Families 1 and 2 have demonstrated relative robustness to the presence of unequal group covariance matrices (see Algina, Oshima, & Tang, 1991), which is reasonable given that their focus is on accounting for this condition by not relying on the pooled covariance matrix, *S*. Furthermore, the performance of these alternatives has proven to be superior to that of the standard Hotelling T^2 when data are multivariate normal but covariance matrices are unequal, both in terms of Type I error rates and power (Holloway & Dunn, 1967). However, these statistics are sensitive to non-normality in the form of moderate to severe skewness (Algina et al., 1991). Coombs, Algina, and Oltman (1996) investigated the Type I error rates of five multivariate generalizations of the Brown-Forsythe and Nel-van der Merwe tests and found that both F_{BF} and F_{NV} were able to maintain the nominal Type I error rate when heterogeneous group covariance matrices were present, but proved to be conservative when the ratio of total sample size to number of dependent variables was small. Christensen and Rencher (1997) observed increases in Type I error rates of F_{JA} and F_{Y}, particularly when the ratio of sample size to number of outcome variables was small. These authors recommended the use of F_{K} for cases in which the group covariance matrices were unequal. However, they acknowledged that this statistic was very conservative for cases in which the sample size to outcomes ratio was between 2 and 3. In a similar fashion, the F_{BF} and F_{NV} tests were shown to be conservative when the assumption of equal covariance matrices was violated and the sample size to outcome variables ratio was small (Coombs, Algina, and Olman, 1996). Additionally, Krishnamoorthy and Xia (2006) reported that F_{MNV} was able to maintain the nominal Type I error rate when group covariance matrices were unequal, as long as the response variables were distributed as multivariate normal. When the latter condition was not met, their test will likely not be appropriate as it relies on multivariate normality. Yanagihara and Yuan (2005) also examined many different versions of modified tests (e.g., *F* statistic, Bartlett correction, modified Bartlett correction) showing that the modified

Bartlett was comparable to the *F* statistic in many cases. This summary of work represents many studies that have examined various test statistics in the MANOVA framework to find a balance between Type I error and statistical power assist in the obtainment of an accurate statistical conclusion.

When the assumption of multivariate normality is violated these parametric MANOVA alternatives exhibit inflated Type I error rates, particularly with small sample sizes (Algina et al, 1991; Fouladi & Yockey, 2002; Wilcox, 1995). Thus, it appears that these alternative statistics are preferable to the standard multivariate test statistics when there are unequal group covariance matrices and the data are normally distributed. However, collectively they do not appear to be robust to violations of multivariate normality, yielding inflated Type I error rates.

Robust alternative test statistics for MANOVA

An alternative approach to the multivariate test statistics when there are violations of the normality assumption involves the use of trimmed means and Winsorized variance (Lix & Keselman, 2004). Statistical problems associated with nonnormality (e.g., Type I error inflation) in the univariate case can be ameliorated by using trimmed means and Winsorized variances in the construction of test statistics (e.g., Lix & Keselman, 2004; Keselman, Kowalchuk, & Lix, 1998; Wilcox, 1995). The use of the trimmed mean involves the removal of the most extreme data points of the response variable in each tail of the observed data distribution. The goal of such a statistic is to avoid the biasing of the mean estimate as a function of one or more outliers in the sample data. Wilcox (1995) recommended censoring 20% of the extreme observations at each tail of the distribution.

The appropriate measure of variation to accompany the trimmed mean is the Winsorized variance (Yuen, 1974). This estimate of variance is based on the Winsorized mean, which is calculated by replacing some portion (e.g., top and bottom 20%) of the most extreme scores in the sample data distribution with the next most extreme scores. The calculation for the Winsorized variance for variable *p* can be seen in Appendix A. As an example of trimming, consider the following set of 10 height measurements in inches: 58, 60, 69, 70, 70, 71, 71, 72, 73, 74. If the recommended 20% trimming were used, a total of 10 x 0.2, or 2, scores are removed. Thus the lower bound value (Y_L) is 60 and the upper bound value (Y_H) is 73, meaning that 58, 60, 73 and 74 are removed from each tail of the distribution, and thereby left out of the calculation of the trimmed mean, which in this case is 70.5. In contrast, the mean based on all 10 observations is 68.8. This is

how trimming was conducted for this study with SAS macros written by Lix and Keselman (2004). In other words, trimming and Winsorizing were conducted along each dimension individually, as described by Lix and Keselman. The Winsorized mean, which will be used in the calculation of the Winsorized variance, is based on 10 data points, with the lowest two values (58 and 60) replaced by 69, and the highest two values (73 and 74) replaced by 72. The value of $\overline{Y}_{wp} =$ 70.5, a 1.7 increase in the value used as the mean.

Lix and Keselman (2004) demonstrated how Winsorized variances and covariances can be applied to multivariate statistics in order to create a Winsorized covariance matrix. Note that the null hypothesis being tested when trimmed means are used involves only the part of the population of interest for which the trimmed mean is appropriate. Thus, the null hypothesis applies to population trimmed means. Given the trimmed means and Winsorized variances for a set of outcome variables, robust alternatives to the test statistics described above can be computed. Specifically, Lix and Keselman (2004) showed that both T^2 and $T^2_{unequal}$ can be calculated using the trimmed means and Winsorized covariance matrices. Likewise, the version of Hotelling's T^2 that does not use the pooled covariance matrix is available. See Appendix A. The robust test statistics will be organized into families using the same logic as described above for their non-trimmed versions; i.e. the trimmed versions reside under their home family (1 or 2).

Prior research regarding robust MANOVA test statistic performance

A number of the MANOVA test statistic alternatives described above based on trimmed means and Winsorized variances have been empirically compared (Wilcox, 1995). Wilcox focused on the case with 4 response variables, with a variety of data distributions, correlations among the response variables and sample sizes. Results showed that when the data were normally distributed, the standard and robust (trimmed) statistics exhibited comparable Type I error rates. However, for non-normal distributions (whether skewed or heavy tailed), the trimmed statistics F_{TK} and F_{TJN} were found to be preferable to their non-trimmed counterparts F_K and F_{JN} in terms of power, and overall, F_{TJN} demonstrated superior control over the Type I error rate for most of the simulated conditions. Beyond Wilcox's (1995) work, there is little empirical work comparing the performance of the robust alternatives to the other alternatives for multivariate mean comparisons when the group covariance matrices are not equal (Lix &

Keselman, 2004). It would appear, therefore, that an extensive evaluation of these methods under a variety of data conditions is warranted. Such work would inform the practitioner of which option may be optimal for use given data conditions. It also is noted that prior comparisons of these methods have been constrained to the two group case.

Rank based nonparametric test

Another alternative approach to dealing with violations of the standard MANOVA assumptions comes in the form of a rank based nonparametric test. A version of this test was first described by Puri and Sen (1971), and then further developed (Erdfelder, 1981; Katz & McSweeney, 1980). The statistic uses the ranks of the raw data as the dependent variables. Erdfelder's extension of this work involves the conversion of the Pillai's trace value obtained from conducting MANOVA using the ranks into the chi-square statistic $\chi^2 = (n-1)P$ (6), where P is Pillai's trace and n is the total sample size. The resulting value is compared with the χ^2 distribution with $k(p-1)$ degrees of freedom, where k is the number of groups for the independent variable and p is the number of response variables as described above. Thus, to compute this rank based nonparametric test, the researcher would first rank each of the dependent variables, and then conduct the MANOVA with the software package of choice, using the ranked dependent variables. The resulting value of P for the independent variable would then be converted using the equation described above. The rank based test represents a third family (Family 3) of statistics considered in this study.

Prior research regarding rank based MANOVA test statistic performance

There has been some empirical evaluation of the performance of the rank based approach, particularly as it compares to the common parametric statistics when the assumptions of normality and/or homogeneity of covariance matrices were violated. Ittenbach, Chayer, Bruininks, Thurlow, and Beirne-Smit (1993), for example, compared the rank based test with the standard MANOVA test statistics and reported somewhat higher power rates for the rank approach. However, Ittenbach and colleagues employed a real dataset for which the population distribution and equality status of the group covariance matrices was not known. Finch (2005) conducted a Monte Carlo simulation study comparing the rank based test statistic with Pillai's trace under a variety of conditions (e.g., normal and non-normal distributions, equal and unequal covariance matrices). When both

assumptions were met, Pillai's Trace and the nonparametric rank test each maintained Type I error rates near the nominal level, but the rank test exhibited lower power. When the assumption of normality was violated, both statistics maintained the nominal Type I error rate of 0.05, regardless of the type of distribution (double exponential, skewed normal, uniform), and had comparable power rates. In the presence of unequal covariance matrices, Finch noted that the rank based nonparametric tests resulted in lower Type I error rates compared to the parametric approach, though both methods had inflated values. Furthermore, as with standard multivariate statistics, the Type I error inflation when there were violations in covariance matrices was more pronounced when group sizes were unequal and the smaller group had the larger variances. Thus, the rank based alternative represents an improvement in the case of unequal covariance matrices, but may not be an ideal solution.

Structural equation models for MANOVA tests

Raykov (2001) suggested the use of structural equation modeling (SEM) as a potential alternative to MANOVA for testing the equality of group mean vectors, particularly when the assumption of equal covariance matrices is violated. He argued that because in the SEM framework covariance matrices can be allowed to differ, this approach might prove superior to the standard MANOVA when group covariances are heterogeneous. This may be an important property, given the aforementioned evidence that other MANOVA test statistics appear to have difficulty in both controlling Type I error and maintaining high power in the heterogeneous covariance case. The basic approach in this case is based on the standard confirmatory factor analysis (CFA) model (see Raykov, 2001), which takes the form:

$$x = \Lambda\xi + \delta$$

where

x = observed variable

ξ = vector of latent variables with covariance matrix Φ

Λ = factor loading matrix

δ = error term

(7)

In most applications of CFA, each latent variable is associated with multiple observed variables. However, in this case each observed dependent variable in the

MANOVA context is related to its own unique latent variable, due to the following strictures:

$$\Lambda = I_p \text{ and } \Theta = 0_{pxp} \tag{8}$$

Here I_p is the identity matrix and Θ is the covariance matrix for δ. In this special case, the covariance matrix for error is comprised of zero elements. These special restrictions, taken together with the CFA model imply that each latent variable is equal to one of the observed variables (Raykov, 2001) and that the latent variable covariance matrix is identical to that of the observed variables. In order to test the null hypothesis of equality of group mean vectors for the response variables, two further assumptions must be made (Raykov, 2001):

$$\begin{aligned} &(1)\ E(\xi) = E(\mu) \\ &(2)\ E(\delta) = 0 \end{aligned} \tag{9}$$

These additional restrictions to the model make the comparison of latent means equivalent to a comparison of observed means. The researcher can then test the null hypothesis of no group difference on the vector of observed dependent variable means by fitting two CFA models, one in which the response variable means are constrained to be equal across groups and the other in which they are allowed to vary. Then, the test of the null hypothesis of group difference on the responses is the difference in the χ^2 fit statistics: $\chi^2_{\text{Constrained}} - \chi^2_{\text{Unconstrained}}$ (10). Allowing the group means to differ results in a saturated CFA model so that the value of $\chi^2_{\text{Constrained}}$ will be 0. Therefore, the test of the null hypothesis of group differences across the vector of dependent variable means is equivalent to $\chi^2_{\text{Constrained}} - \chi^2_{\text{Unconstrained}} = \chi^2_{\text{Constrained}} - 0 = \chi^2_{\text{Constrained}}$ (11; Raykov, 2001).

As noted above, the primary advantage of using the SEM approach to compare group mean vectors is that covariance matrices can be allowed to vary across groups (Raykov, 2001). In this way, the assumption of covariance matrix equality which underlies standard MANOVA and which has been shown in prior research to be important for other statistics for testing multivariate mean equality, is no longer a requirement. When the assumption of normality is violated, the standard χ^2 statistic used with ML estimation in SEM may not perform well (Yu & Muthén, 2002). Therefore, an adjusted version of this test statistic is appropriate when the dependent variables are not normally distributed. This alternative, developed by Satorra and Bentler (1994), was designed to correct for

multivariate kurtosis, and has been shown to be robust to departures from multivariate normality (Curran, West, & Finch, 1996).

Given that the MANOVA test statistics are not as accurate as desired under violations of assumptions, alternative methods need to be explored to test the same hypotheses (Raykov, 2001) yet such evaluations have not occurred in sufficient number. As described above, prior simulation research examining alternatives to the standard MANOVA approach for testing multivariate mean differences (e.g., rank based and exact tests) has generally found that assumption violations, particularly that of homogeneity of covariance matrices, result in Type I error inflation similar to, if not as severe as, that reported for MANOVA (e.g., Finch, 2005; Ittenbach, Chayer, Bruininks, Thurlow, & Beirne-Smith, 1993). By contrast, very little empirical research has been conducted to evaluate the effectiveness of this fairly new SEM based approach for testing the null hypothesis of multivariate mean equality. One such effort (Finch & French, 2008) found that in the absence of assumption violations, the Satorra-Bentler corrected χ^2 test and Pillai's trace had comparable Type I error rates and power for total samples of 100 or more with normally distributed dependent variables. For smaller samples, the SEM based approach did have elevated Type I error rates (e.g., 0.09 for *N* of 30) when both assumptions of normality and homogeneity of covariance matrices were met. When the assumption of equal covariance matrices was violated and the smaller group had the larger elements, the SEM based approach had lower Type I error rates compared to the standard approach. When the larger group had the larger elements, both SEM and the standard approach had Type I error rates at or below the nominal level, but the SEM method had much higher power. Thus, it appeared that the SEM approach might be preferred. However, there is a need to examine the large number of viable MANOVA test statistics reviewed here under the same conditions to begin to inform the field as to which approach is preferred under different conditions. Additionally, little, if any prior work has examined the performance of this new SEM approach to MANOVA testing as well as with more than two groups. The SEM approach to testing hypotheses about multivariate mean differences represents a fourth family (Family 4) of test statistics investigated in this study.

Goals of the study

The first goal of this study was to review the various MANOVA test statistics to inform the reader of the 16 choices that are currently available for comparing multivariate means across groups. Table 1 provides summary information across

these 16 tests to aid understanding of performance from separate past evaluations. The second goal was to conduct a simulation study comparing the performance of the 16 methods across a variety of conditions designed to mirror those encountered in practice, in order to assess their Type I error and power rates. This Monte Carlo study is anticipated to provide information on performance of these tests to aid the researcher in selecting the test that appears to work well given the specific data at hand and corresponding assumptions that are or are not met. The literature review led to several predictions for comparing test statistics noting that it is impractical to make predictions for all combinations investigated. First, it was expected that when the data were normally distributed and group covariance matrices were homogeneous, all methods would have comparable Type I error and power rates. Second, Families 1, 2, 3 and 4 were expected to have, on average, lower Type I error and higher power compared to the standard MANOVA test statistic, when covariance matrices were heterogeneous. Third, given the advantages of latent variable modeling it was expected that SEM would have the lowest Type I error and highest power, across conditions, with the exception of for small sample sizes, where accurate parameter estimation would likely be a problem. Fourth and last, trimmed versions in Family 1 were expected to have the lowest Type I error and highest power in heavily skewed distribution conditions.

A number of studies have previously conducted investigations of a few of these methods, but no study has simultaneously compared all of the techniques under a common set of conditions. In addition to all comparisons under similar conditions, this work adds to the literature by providing information on the use of SEM under these conditions and behavior of all statistics studied for the 2 and 3 group case. The former is rarely included in such comparisons and no evaluation has investigated performance of all four test families in one simulation under the same conditions. Thus, the present work seeks to extend the literature by providing a full examination of methods for comparing multivariate group means when standard assumptions are not met. A total of seven factors were manipulated which allowed for the examination of 12,076 conditions to assist with meeting the second goal of the study.

Methods

This Monte Carlo study manipulated seven factors in a completely crossed design with 5000 replications per combination of conditions using a SAS program (SAS version 9.1, 2004) written by the authors. Manipulated factors included sample size, group size ratio, covariance matrix homogeneity/heterogeneity, distribution

of the dependent variables, group mean differences, correlations among the dependent variables, and the number of dependent variables. All of the statistical methods were conducted using SAS, with the exception of SEM, which was carried out with M*plus* version 5.1 (Muthén & Muthén, 2008). A number of the alternative and robust methods were conducted with a macro described by Lix and Keselman (2004). The two outcome variables of interest were the Type I error rate (rejecting the null hypothesis of no multivariate mean difference when, actually, no differences were simulated) and power (correctly rejecting the null hypothesis of no multivariate mean differences). To assess which of the manipulated factors, or combinations of them, had a significant influence on the dependent variables, an ANOVA was conducted for each outcome, per recommendations for simulation research (Paxton, Curran, Bollen, Kirby, & Chen, 2001). The dependent variable in each ANOVA was one of the outcomes (i.e., Type I error rate or power) taken across replications for each combination of conditions. The independent variables were the manipulated factors and their interactions. In addition, the ω^2 effect size was calculated to describe the relative magnitude of the statistically significant effects. Given the scope of the simulation, discussion is limited only to those effects that most influenced the Type I error and power rates, which are defined as those effects that were both statistically significant ($\alpha = 0.05$) and had an ω^2 of 0.10 or greater.

Statistical methods

Because it has been demonstrated as more robust with respect to Type I error control when standard assumptions are violated (Olson, 1974), *Pillai's Trace* (*P*) was selected for use as the standard MANOVA test statistic for this study, and will be referred to as such throughout the remainder of the manuscript, although it is acknowledged that other test statistics such as Wilks' Lambda, are also frequently used in practice. However, note that with the two groups case the results across the standard tests will be identical, and equal to Hotelling's T^2. The other statistical tests included the rank based method, James (JA), Hotelling's T^2 for unequal covariance matrices (H), Brown-Forsythe (BF), Johansen (JO), Kim (K), Nel van der Merwe (NV), Yao (Y), Raykov (SEM), and the trimmed versions of the robust methods, TJA, TH, TBF, TJO, TK, TNV, and TY. Consistent with the recommendation of Lix and Keselman (2004), 20% symmetric trimming of the data was employed.

Manipulated Factors

Total sample size

Seven total sample size (across groups) conditions were examined for the two groups case: 20, 30, 45, 60, 90, 100, and 150. For the three groups case, the following total sample size conditions were examined: 30, 40, 45, 50, 60, 75, 90, 120, 150, 200, and 250. In the three groups, equal sample size condition for N=40, 50, 200, and 250, the data were simulated so that one group had either one more or one fewer observations than did the others. For example, in the N=40 case, two of the groups were simulated with 16 individuals, whereas the other was simulated with 17. Similarly, in the N=250 condition, two of the groups were simulated with 83 individuals, whereas the other was simulated with 84. The same approach was used with 50 and 200. These values are in accord with previous simulation research with MANOVA and SEM approaches to multivariate comparisons, (e.g., Christensen & Rencher, 1997; Finch, 2005; Hancock, Lawrence & Nevitt, 2001; Hussein & Carriere, 2005; Wilcox, 1995). This range of values was selected to reflect conditions that applied social science researchers are very likely to encounter.

Number of Groups

Two conditions were simulated for number of groups: 2 and 3 groups. Much of the previous work comparing performance in the MANOVA situation has been conducted on 2 groups with several variables (e.g. Christensen & Rencher, 1997; Finch, 2005). A significant addition of this work to the literature is to evaluate the behavior of these tests with 3 groups. Two groups were included to aid the comparison to prior work.

Group size ratio

Three group size ratio conditions were used: (a) groups were equal, (b) group 1 was half the size of group 2, and (c) group 1 was twice the size of group 2. In the three group case, for condition (b) groups 1 and 2 were half the size of group 3, and for condition (c) groups 1 and 2 were twice the size of group 3. Thus, for example, in the n=60 case, there were 30 simulees per group in condition a, 20 in group 1 and 40 in group 2 in condition b, and 40 in group 1 and 20 in group 2 in condition c. The combination of unequal group sizes with unequal group covariance matrices has been shown to influence both Type I error and power

rates (Sheehan-Holt, 1998; Stevens, 2001; Hakstian, Roed & Lind, 1979) and these particular ratios employed have been used in prior studies (e.g., Christensen & Rencher, 1997, Hakstian et al.,1979). As noted above, when the smaller group has the larger covariance matrix elements the Type I error rate will be inflated; when the larger group has the larger elements power will be diminished.

Covariance matrix homogeneity/heterogeneity

The group covariance elements were manipulated in three ways: (a) equal, (b) one group had elements 5 times as large as the others, and (c) one group had elements 10 times as large as the others. Equality of the covariance matrices across groups is a vital assumption for the test statistics associated with MANOVA, and differences in these matrices can influence the performance of these tests (Fouladi & Yockey, 2002; Sheehan-Holt, 1998; Korin, 1972). Two unequal covariance conditions were simulated (a) the larger group had the larger elemental values and (b) the smaller group had the larger elemental values.

Distribution of the dependent variables

Normality of the dependent variables is another assumption of the standard statistical tests used in MANOVA. The Type I error rate of the common MANOVA tests may suffer from some inflation when the distribution of the dependent variables have large kurtosis (Olson, 1974). Therefore, in the current research the dependent variables were simulated under one of four distributional conditions: (a) normal (skewness=0, kurtosis=0), (b) beta (skewness = -0.82, kurtosis = 0.28), (c) lognormal (skewness = 6.18, kurtosis = 110.93), and (d) exponential (skewness =2, kurtosis = 6). These reflect conditions used in similar work (Algina et al., 1991). The non-normal data were simulated using a methodology described by Headrick and Sawilowsky (1999) to achieve the desired levels of skewness and kurtosis while maintaining the target correlations among the dependent variables. These distributions were selected to provide insights into the performance of the methods studied here under a variety of cases.

Group means differences

Differences in group means were simulated using values of Cohen's (1988) *d* univariate effect sizes. This metric was selected because it allowed for a straightforward manipulation of this important variable and matches the methodology (though not the values) used in prior simulation research of MANOVA (Blair et al., 1994; Finch, 2005). The effect size of 0 allowed for the evaluation of the Type I error. The other values corresponded to group separation

at small (0,2), medium (0.5), and large (0.8) levels. The univariate Cohen's *d* (i.e., $mean_{group1} - mean_{group2} / SD_{pooled}$) effect size was selected for use in this study because there are generally agreed upon guidelines for its interpretation (Kim & Olejnik, 2004). In contrast, though there do exist multivariate effect size values, there is not a single such statistic that is considered the standard, nor is there any sort of agreement regarding what constitutes a small, medium, or large effect. Thus, in order to provide a useful context to researchers regarding the performance of the various methods described here, Cohen's *d* was used.

Correlation among the dependent variables

The data were simulated under three conditions for correlation among the dependent variables, including no correlation (0.0), small (0.2) and large (0.8). These values were selected to represent the case where variables were orthogonal (0.0), where the correlation was small to moderate (0.2) and where the variables were highly correlated (0.8). Conditions are consistent with prior research (e.g., Finch, 2005; Wilcox, 1995) to aid comparability.

Number of dependent variables

Two levels were employed: 2 and 4 dependent variables, consistent with prior studies (e.g., Fouladi & Yockey, 2002; Wilcox, 1995) and representative of realistic numbers of response variables seen in practice (e.g., Dumas et al., 1999; Krull, Kirk, Prusick, & French, 2010) while maintaining a manageable set of simulation conditions for the current study.

Results

Two groups versus three groups

Results for two and three group cases generally followed similar patterns in terms of how the methods compared relative to one another, with a couple of exceptions. Thus, to keep discussion of the results as brief as possible, only results for the three group condition are presented. However, prior to presenting these, note that the few cases where the two group condition results diverged from those for three group condition. In general, Type I error rates did not differ between the two number of groups conditions, but power was higher in the three group case compared to the two group case. In terms of relative comparison of the methods, with two groups the rank based approach had among the lowest power values. When three groups were present, the rank based approach had comparable power

to the other approaches, as is presented below. Outside of these differences, the results for the two group case were comparable to those for three group case, which are presented below. The two group case results are available from the authors upon request.

The results for three group case are organized into two sections: (a) Type I error and power rates based on the variance homogeneity condition, and (b) Type I and power rates by the distribution of the response variables. In each case, a repeated measures ANOVA was employed to identify the significant main effects and interactions of the manipulated factors in terms of Type I error and power, where the repeated measures variable was the MANOVA test statistic. The ANOVA models had as the dependent variable the Type I error or power rates across the 5000 replications per combination of conditions. The independent variables were type of test statistic (within replication), correlation among the dependent variables, number of dependent variables, sample size ratio, variance ratio, sample size and in the case of power, and effect size. The assumptions of normality and sphericity were assessed and found to have been met. Sphericity was assessed using Mauchley's test of Sphericity in conjunction with the ε statistic, which takes the value of 1 in the population when the covariance matrix satisfies sphericity (Warner, 2008). In the case of each set of repeated measures ANOVA results below, Mauchly's test was not statistically significant with α = 0.15, as recommended in Kirk (1995). In addition, across the repeated measures analyses described below, the value of ε ranged between 0.901 and 0.974. Finally, an examination of the Greenhouse-Geisser conservative *F*-test and MANOVA test results, both of which have been suggested for use when sphericity is violated, revealed the same main effects and interactions as significant and non-significant when compared with the unadjusted test. Therefore, given the general finding that sphericity was present, coupled with the similarity in results for the unadjusted and Greenhouse-Geisser adjusted test, it may be concluded that sphericity (or lack thereof) was not problematic in this case.

Normality was assessed first by an examination of QQ-plots for the individual outcome variables, and all were found to conform reasonably closely to the line for the normal distribution. In addition, multivariate normality was tested for across repeated measurements (rejection rates for each test statistic) for each of the models described below using Mardia's test (Mardia, 1970), and found none of them to be statistically significant. Taken together, the QQ-plot and Mardia's test results satisfy the assumption of normality for repeated measures models as described in Warner (2008). The models were fully factorial with all main effects and interactions included. As mentioned previously, in order to focus

on only the most important of the manipulated factors, discussion will be limited to those significant (α = 0.05) terms in the ANOVA that had an effect size (ω^2) greater than 0.10. This value was selected because it indicates that the main effect or interaction term accounted for at least 10% of the variation in rejection rates. By doing so, it is possible to avoid discussing a large number of statistically significant effects that actually accounted for a small amount of variance, which was a concern given the large number of replications for each combination of conditions. Full results tables are available by contacting the authors.

Covariance Homogeneity: Type I error rate

The ANOVA identified the interaction of test statistic by sample size ratio by covariance ratio as the highest order significant term ($p < 0.01$, $\omega^2 = 0.527$). The interaction of test statistic by number of dependent variables was also significant ($p < 0.01$, $\omega^2 = 0.381$). No other term was statistically significant with an effect size value greater than 0.10.

Table 2 contains the Type I error rates by test statistic, sample size ratio, and covariance ratio for normally distributed data. When the groups' covariances were equal, the Type I error rate for all of the statistics examined here were below 0.06, except for H, TH, and the rank approach across group ratio conditions, and for BF in the sample size ratio 2/1 condition. When the group covariances were not equal but the sample size ratio was equal, the Type I error rate of the *P* test was inflated above the nominal 0.05 level. Several of the alternative statistics, including the rank based approach, H, TH, and BF had inflated Type I error rates in the unequal covariance, equal sample size condition as well. In contrast, the Type I error rates for JA, JO, K, NV, Y, all members of Family 1 (except for K), and SEM did not have inflated error rates associated with inequality in group covariances when sample sizes were equal. To further investigate these effects, several interaction contrasts were employed, using Scheffé's correction (Scheffé, 1953) to control the overall Type I error rate (α = 0.05) and allow for such post hoc investigations. Based on these contrasts, it was found that the rank and H statistics yielded significantly inflated Type I error rates as the degree of covariance inequality increased, whereas the rates of the other methods did not change significantly.

Table 2. Type I Error Rate by Test Statistic, Sample Size Ratio, and Group Covariance Ratio: Normally Distributed Data

Sample Size Ratio	Statistic	Covariance ratio: 1/1	Covariance ratio: 5/1	Covariance ratio: 10/1
Equal	*Standard*	0.050	0.060	0.064
	Ranks	0.060	0.072	0.082
	JA | *TJA*	0.043 | 0.031	0.049 | 0.033	0.052 | 0.040
	H | *TH*	0.070 | 0.070	0.089 | 0.093	0.102 | 0.117
	BF | *TBF*	0.053 | 0.038	0.071 | 0.046	0.072 | 0.046
	JO | *TJO*	0.051 | 0.042	0.056 | 0.044	0.056 | 0.050
	K | *TK*	0.047 | 0.034	0.040 | 0.027	0.035 | 0.024
	NV | *TNV*	0.048 | 0.034	0.047 | 0.029	0.047 | 0.028
	Y | *TY*	0.048 | 0.035	0.055 | 0.040	0.059 | 0.044
	SEM	0.057	0.054	0.058
1/2*	*Standard*	0.050	0.007	0.004
	Ranks	0.061	0.023	0.019
	JA | *TJA*	0.044 | 0.033	0.048 | 0.042	0.046 | 0.040
	H | *TH*	0.061 | 0.061	0.031 | 0.034	0.024 | 0.018
	BF | *TBF*	0.052 | 0.043	0.064 | 0.075	0.067 | 0.081
	JO | *TJO*	0.051 | 0.045	0.051 | 0.041	0.046 | 0.040
	K | *TK*	0.051 | 0.042	0.046 | 0.039	0.042 | 0.036
	NV | *TNV*	0.047 | 0.035	0.047 | 0.042	0.044 | 0.041
	Y | *TY*	0.053 | 0.046	0.048 | 0.043	0.049 | 0.039
	SEM	0.053	0.055	0.061
2/1**	*Standard*	0.049	0.092	0.109
	Ranks	0.061	0.086	0.103
	JA | *TJA*	0.042 | 0.033	0.047 | 0.037	0.050 | 0.041
	H | *TH*	0.068 | 0.065	0.122 | 0.121	0.157 | 0.159
	BF | *TBF*	0.064 | 0.052	0.069 | 0.050	0.070 | 0.049
	JO | *TJO*	0.053 | 0.048	0.055 | 0.046	0.055 | 0.048
	K | *TK*	0.050 | 0.041	0.041 | 0.033	0.037 | 0.029
	NV | *TNV*	0.044 | 0.039	0.047 | 0.033	0.048 | 0.033
	Y | *TY*	0.058 | 0.047	0.055 | 0.047	0.055 | 0.046
	SEM	0.049	0.055	0.053

*Sample size ratio of 1/2 couples larger variance with larger group size in the unequal variance condition.
**Sample size ratio of 2/1 couples larger variance with smaller group size in the unequal variance condition.

Based on interaction contrasts using Scheffé's corrected critical value, when the larger group had the larger covariance (sample size ratio 1/2), the *P*, rank based statistic and H displayed significant declines in Type I error rates concomitant with increases in groups' covariance matrix inequality. As the group

covariances became more unequal, however, TBF had a significant increase in the Type I error rate. As seen with an equal sample size ratio, members of Family 1 and K generally demonstrated consistent Type I error rates, which were just below the nominal value of 0.05. The Scheffé corrected contrasts did not find any significant change in the error rates of the SEM method, though for the covariance ratio of 10/1 with the 1/2 sample size ratio, the rate was just above 0.06. When the smaller group had the larger covariance (sample size ratio 2/1), the standard, rank based, H, and TH approaches all showed a significant increase in the Type I error rate with increasing divergence in group covariance matrices. Family 1 and SEM maintained Type I error rates near the nominal 0.05 value, whereas K actually had a slight decline in the error rate as the covariance matrices became more unequal. Across all conditions simulated here, the trimmed versions of the test statistics had slightly lower Type I error rates compared to the untrimmed alternatives (except for TH in the covariance ratio 10/1, sample size ratio 2/1 case), though in most cases these differences were less than 0.01.

Table 3. Type I Error Rate by Test Statistic and Number of Dependent Variables: Normally Distributed Data

	Number of dependent variables	
Statistic	**2**	**4**
Standard	0.069	0.087
Ranks	0.065	0.079
JA | *TJA*	0.043 | 0.042	0.041 | 0.039
H | *TH*	0.072 | 0.074	0.074 | 0.078
BF | *TBF*	0.062 | 0.052	0.064 | 0.049
JO | *TJO*	0.049 | 0.044	0.051 | 0.042
K | *TK*	0.048 | 0.041	0.043 | 0.040
NV | *TNV*	0.050 | 0.046	0.052 | 0.039
Y | *TY*	0.046 | 0.046	0.053 | 0.038
SEM	0.057	0.051

Table 3 displays the Type I error rate for statistical test by number of dependent variables for normally distributed data. The error rates for the standard and rank based approaches were significantly greater for 4 variables compared to 2 variables. The Type I error rates for the rest of the test statistics were essentially the same for 2 and 4 dependent variables. In addition to the standard and rank approaches, H, TH, and BF all had error rates in excess of 0.06; the other methods

had rates closer to the nominal 0.05. Because there were not significant results for the correlation among the dependent variables and the sample size, they are not discussed.

Covariance Homogeneity: Power

Repeated measures ANOVA was used to identify the manipulated terms that were significantly related to power rates across replications, using the same model used with Type I error rates. The interaction of the test statistic by sample size ratio by covariance ratio was the highest order significant term ($p < 0.01$, $\omega^2 = 0.149$), as were the main effects of effect size ($p < 0.01$, $\omega^2 = 0.811$), correlation among the dependent variables ($p < 0.01$, $\omega^2 = 0.360$) and total sample size ($p < 0.01$, $\omega^2 = 0.781$). No other terms in the ANOVA were statistically significant with an effect size greater than 0.10.

Table 4 contains power by test statistic, sample size ratio and group covariance ratio. Power values for those conditions for which the Type I error rate was greater than 0.075 (from Table 2) are in bold, and should be interpreted with extreme caution. These values are included for completeness in results presentation. When the groups were of equal size, SEM, followed by the *P* statistic had the highest power rates among those for which the Type I error rates were not inflated (non-bolded values). For all of the methods studied here, power declined as the covariance matrix inequality increased when the larger group had the larger variance and when the smaller group had the larger variance. In addition, the power for the trimmed statistics was uniformly lower than that of the non-trimmed versions in this sample. Power for the rank based approach was comparable to that of the standard in the covariance 1/1 and 5/1 cases, but could not be interpreted for 10/1 due to Type I error inflation.

When the group sizes were unequal but the covariance matrices were equal, SEM had the highest power rates, followed by the standard, and rank based approaches, all of which had significantly higher power than the other methods studied here. When the larger group had the larger covariance (sample size 1/2 condition), power for all methods declined significantly with increases in variance heterogeneity, though the pattern of SEM, followed by standard and rank methods with highest power rates held. When the smaller group had the larger covariance (sample size 2/1 condition), a situation that resulted in inflated Type I error rates for several methods, the highest power rates among those that had Type I error rates lower than 0.075 belonged to SEM, followed by Family 1, K, BF, and TBF. For all of the methods power rates declined significantly as the degree of

covariance matrix inequality increased. Note that in this condition, the Type I error rates for the standard, rank based, and H approaches were inflated.

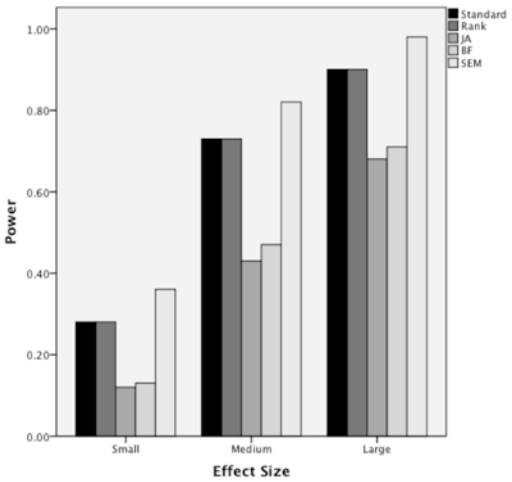
Table 4. Power by Test Statistic, Sample Size Ratio, and Group Covariance Ratio: Normally Distributed Data

Sample Size Ratio	Statistic	Covariance ratio: 1/1	Covariance ratio: 5/1	Covariance ratio: 10/1
Equal	Standard	0.695	0.44	0.309
	Ranks	0.684	0.464	**0.353**
	JA | TJA	0.470 | 0.394	0.256 | 0.203	0.170 | 0.131
	H | TH	0.530 | 0.496	**0.330 | 0.309**	**0.248 | 0.241**
	BF | TBF	0.495 | 0.421	0.302 | 0.230	0.209 | 0.148
	JO | TJO	0.490 | 0.432	0.268 | 0.223	0.178 | 0.145
	K | TK	0.480 | 0.402	0.240 | 0.190	0.142 | 0.102
	NV | TNV	0.483 | 0.405	0.253 | 0.189	0.162 | 0.112
	Y | TY	0.481 | 0.404	0.266 | 0.210	0.177 | 0.135
	SEM	0.738	0.489	0.357
1/2*	Standard	0.764	0.538	0.413
	Ranks	0.758	0.55	**0.435**
	JA | TJA	0.537 | 0.463	0.319 | 0.267	0.222 | 0.184
	H | TH	0.587 | 0.558	**0.389 | 0.369**	**0.312 | 0.300**
	BF | TBF	0.558 | 0.500	0.363 **| 0.300**	0.261 **| 0.206**
	JO | TJO	0.558 | 0.506	0.332 | 0.288	0.230 | 0.194
	K | TK	0.551 | 0.491	0.310 | 0.259	0.201 | 0.159
	NV | TNV	0.545 | 0.468	0.321 | 0.261	0.220 | 0.171
	Y | TY	0.557 | 0.499	0.332 | 0.283	0.229 | 0.188
	SEM	0.802	0.591	0.472
2/1**	Standard	0.741	**0.705**	**0.685**
	Ranks	0.734	**0.721**	**0.69**
	JA | TJA	0.514 | 0.440	0.498 | 0.403	0.377 | 0.334
	H | TH	0.568 | 0.537	**0.536 | 0.511**	**0.436 | 0.367**
	BF | TBF	0.537 | 0.473	0.492 | 0.397	0.381 | 0.326
	JO | TJO	0.535 | 0.482	0.488 | 0.401	0.379 | 0.319
	K | TK	0.527 | 0.461	0.487 | 0.410	0.384 | 0.322
	NV | TNV	0.525 | 0.447	0.500 | 0.389	0.380 | 0.343
	Y | TY	0.532 | 0.467	0.519 | 0.402	0.399 | 0.338
	SEM	0.811	0.732	0.691

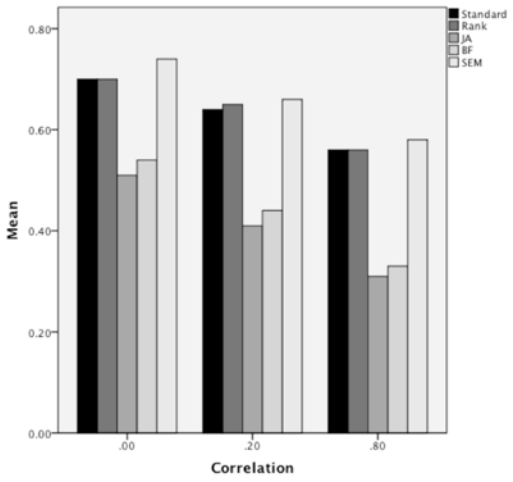
Note: Bold indicates when power values for these conditions were associated with Type I error rates greater than 0.075

*Sample size ratio of 1/2 couples larger variance with larger group size in the unequal variance condition.

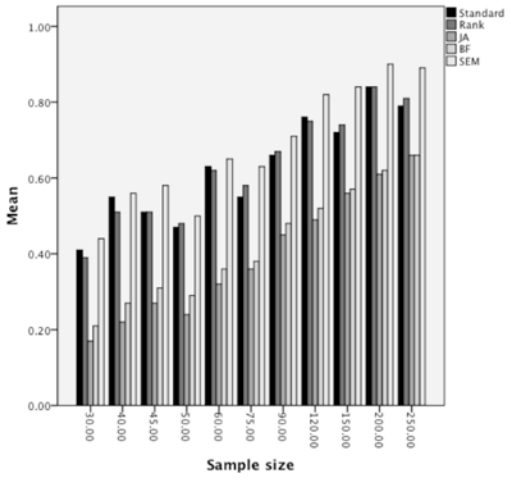
**Sample size ratio of 2/1 couples larger variance with smaller group size in the unequal variance condition.



Power by effect size



Power by correlation among dependent variables



Power by sample size

Figure 1. Three panels of power rates for five MANOVA statistics by effect size, correlation among dependent variables, and sample size.

Figure 1 displays power by the main effects of effect size, correlation among the dependent variables and total sample size, in three panels. For clarity of presentation only selected testing methods were included, as they are representative of others studied. Specifically, JA was selected to represent Family 1 (except H) and K, all of which had very similar rates, though BF displayed similar power to H under these conditions. The trimmed versions of these statistics had power rates that were similar to the untrimmed versions in terms of their pattern relative to one another and had slightly lower power values (though not significantly lower) than the untrimmed statistics. For all of the methods,

power increased significantly with increases in effect size and sample size, and declined with increases in the correlations among the dependent variables. These patterns were consistent across the methods studied here.

Distribution: Type I error rate

As with the covariance homogeneity data, a fully factorial repeated measures ANOVA was used to identify significant main effects and interactions of the manipulated variables that were related to the Type I error rates under differing distribution conditions. The highest order term that was identified as statistically significant with ω^2 greater than 0.10 was the interaction of type of test statistic (method) by number of dependent variables by sample size ($p < 0.01$, $\omega^2 = 0.624$). In addition, the distribution of the dependent variables was a significant main effect ($p = 0.034$, $\omega^2 = 0.063$). Although its ω^2 value did not meet the 0.10 threshold used to identify terms for further consideration, it will be discussed briefly because the distribution of the response was of primary interest in this study. No other term was both statistically significant in the ANOVA and had ω^2 greater than 0.10.

Table 5. Type I Error Rate by Test Statistic and Distribution of the Dependent Variables.

	Distribution			
Statistic	**Normal**	**Beta**	**Lognormal**	**Exponential**
Standard	0.05	0.05	0.05	0.05
Ranks	0.079	0.06	0.061	0.06
JA | TJA	0.047 | 0.036	0.044 | 0.032	0.044 | 0.033	0.044 | 0.032
H | TH	0.104 | 0.106	0.064 | 0.065	0.064 | 0.065	0.064 | 0.064
BF | TBF	0.064 | 0.046	0.052 | 0.042	0.052 | 0.042	0.052 | 0.041
JO | TJO	0.054 | 0.046	0.051 | 0.044	0.051 | 0.045	0.051 | 0.044
K | TK	0.042 | 0.032	0.049 | 0.039	0.048 | 0.040	0.048 | 0.039
NV | TNV	0.047 | 0.032	0.047 | 0.035	0.047 | 0.036	0.047 | 0.035
Y | TY	0.054 | 0.044	0.052 | 0.043	0.051 | 0.043	0.051 | 0.042
SEM	0.055	0.082	0.084	0.082

Table 5 contains the Type I error rate for the test statistics by the distribution of the dependent variables. These results demonstrate that the *P* test statistic was robust to the distribution of the dependent variables, maintaining the nominal (0.05) Type I error rate across the four distributions. With the exception of the

rank based approach, BF, H, and TH in the normal case, and ranks, H/TH, and SEM in the nonnormal conditions, which had elevated rates, the tests displayed Type I error at the nominal level of 0.05.

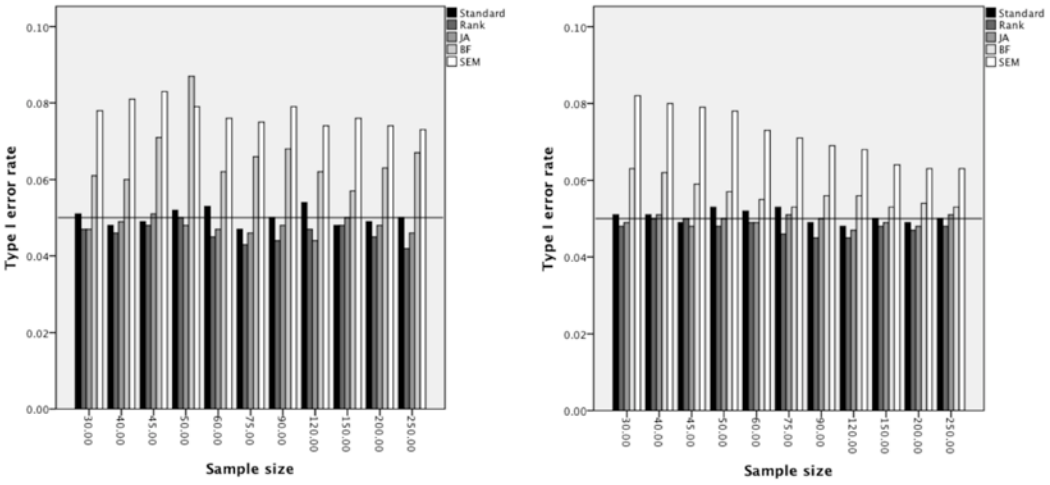


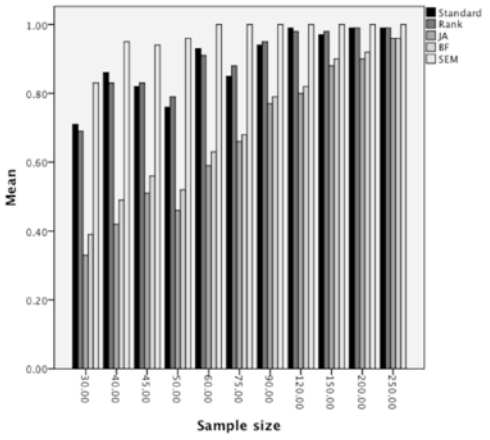
Figure 2. Two panels of Type I error rates for five MANOVA tests by sample size and number of dependent variables, across distribution of the dependent variables.

Figure 2 contains two panels showing the Type I error rates for the methods by the number of dependent variables and the sample size, across distribution conditions. In order to simplify presentation of the results, only the selected methods described were examined, which are representative of other several others that performed extremely similarly. An examination of Figure 2, which has a reference line at the nominal α rate of 0.05, reveals that when there were 2 dependent variables, BF and SEM consistently had elevated Type I error rates. The other methods largely maintained the nominal rate across sample sizes, although the standard statistic did have slightly rates slightly above the 0.05 line (though not as high as 0.06) at N=60 and 120. With 4 dependent variables the standard, rank, and JA methods exhibited Type I error rates near or just below the 0.05 level, except for the standard statistic with samples of 50, 60, and 75, with rates slightly above the nominal rate but not breaking 0.06. In contrast, the error rates for SEM and BF were consistently elevated above 0.05, but declined with increasing sample size. SEM had the highest rates compared to any method. Please note again that these results combine the outcomes for all of the

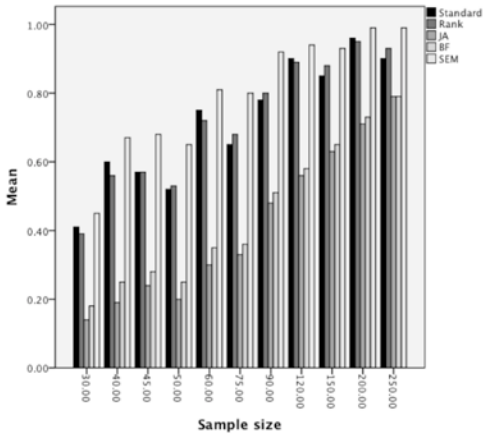
distributions, and that SEM maintained the nominal Type I error rate when the data were normally distributed, though it did not for the nonnormal data.

Distribution: Power

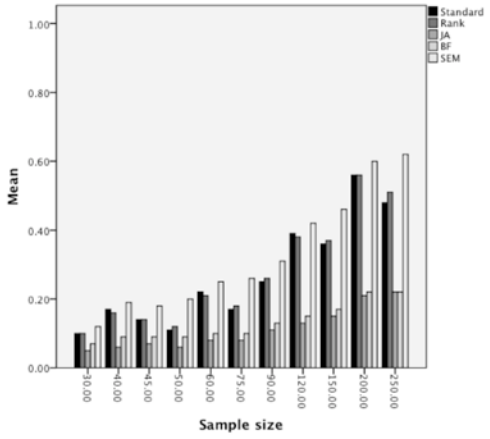
The factorial repeated ANOVA for the power of the MANOVA test statistics when the distributions were varied identified the interaction of method by correlation among dependent variables by distribution by number of variables ($p < 0.001$, $\omega^2 = 0.588$) and the interaction of method by sample size by effect size ($p < 0.001$, $\omega^2 = 0.694$) as the highest order significant terms with ω^2 greater than 0.10. All other significant lower order main effects and interactions were subsumed in these interactions and will not be discussed further.



Power for Large effect size



Power for Medium effect size



Power for Small effect size

Figure 3. Three panels of power rates for five MANOVA tests by effect size and sample size.

Figure 3 (three panels) displays power for the representative statistics used previously (standard, rank based, JA, BF and SEM) by effect size and sample size. When interpreting these results, it is important to keep in mind that the Type I error rates for SEM were inflated when the data were not normally distributed, and therefore higher power rates with SEM must be viewed with caution. The following discussion will focus on power for those statistics that maintained the nominal Type I error rate of 0.05. Across effect size and sample size values, the standard and rank based approaches maintained the highest power values of those methods that were able to maintain the nominal Type I error rate across distributions. In contrast, when the effect size was large, the BF and JA methods had lower power compared to the other approaches for the smallest sample size condition. Not until $N = 120$ did power approach 0.8 for these methods. When the simulated effect size was of medium magnitude, none of the methods that controlled Type I error had power rates approaching 0.8 until sample sizes were 90, and again the standard and rank approaches had higher power than JA or BF. In contrast, for the large effect condition the standard and rank statistics had markedly higher power rates across sample sizes, with values of 0.8 or greater for N of 60 or more. Finally, when the simulated effect size was small, the patterns were similar to those for larger effects, though none of the methods that controlled Type I error had power greater than 0.6 for any sample size, and the standard and rank based approaches had higher power than JA or BF.

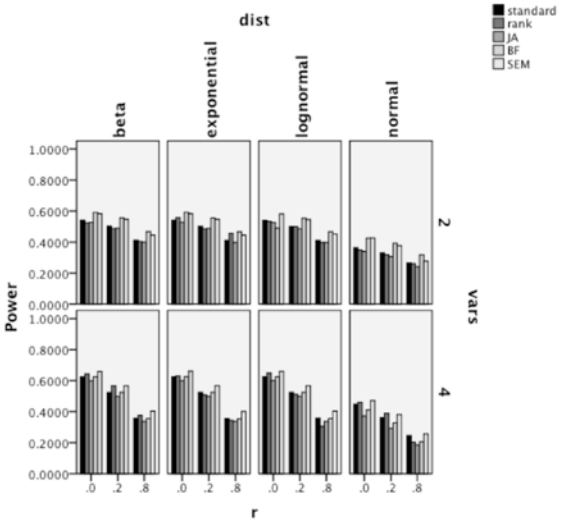


Figure 4. Power for five MANOVA tests by correlation (r) among dependent variables, distribution (dist) of dependent variables, and number of dependent variables (var)

Figure 4 includes the power rates for the significant 4-way interaction of test statistic by correlation among the dependent variables by distribution of the dependent variables by number of dependent variables. An examination of this figure reveals that across distributions and test statistics, power declined with increases in the correlation among the dependent variables. In terms of the test statistics that controlled Type I error, power for BF was generally the highest in the 2 variable case, and the standard, rank, and BF approaches displayed comparable power with 4 variables across distributions. With the normal distribution, SEM also had among the highest power values in the 4 variable condition, on par with standard and rank tests. And again, although SEM had the highest power values in most of the nonnormal conditions, it is not discussed in that context here due to the Type I error inflation it exhibited for the nonnormal distributions. JA consistently displayed among the lowest power results of the methods studied here. Power was consistently lower in the normal distribution condition compared to the other distributions studied here. Finally, note that power was below 0.80 across all conditions.

Discussion

The goals of this study were to provide a comprehensive review of the various test statistics available for MANOVA when standard assumptions are violated, and to conduct a large simulation study to compare the performance of the 16 identified (i.e., four families) test statistics across a variety of simulated conditions to evaluate Type I error and power. The results illustrate that Type I error and power do differ based on the selection of the test statistic for the MANOVA, dependent upon specific data conditions. This work is in accord with calls to make such comparisons. Raykov (2001), for example, encouraged comparison of the standard approach to testing the multivariate null hypothesis of no mean vector difference across groups as represented by *P* with an approach based upon SEM. This comparison was made, among several others, and extended this work to the 3 group case. Thus, this study does provide information on performance of these tests to aid the researcher in selecting the test statistic(s) that appears to work well given the data at hand, corresponding assumptions that are satisfied, and the variable framework (latent vs. observed) under which the analysis is conducted. Seven factors were manipulated resulting in 12,076 comparison conditions to gain a greater understanding of the relative performance of the standard approach for testing the multivariate hypotheses with respect to mean differences, along with a number of purportedly more robust options.

Major Points

Results revealed that when MANOVA assumptions are met, SEM and *P* are optimal in terms of Type I error and power rates. This result for *P* is consistent with prior research (e.g. Christensen & Rencher, 1997), though there is not a great deal of prior work examining many of the other alternative test statistics. Furthermore, both SEM and *P* maintained the nominal error rate in this condition, and SEM had the highest power rates. Even when data are not normally distributed, the *P* statistics maintain the nominal Type I error rate as do the Family 1 and Family 2 test statistics, thus partially supporting the first research hypothesis for this study. However, when the assumption of equal covariances is violated, but group sample sizes remains equal, the *P* statistic displays elevated Type I error rates whereas both SEM and Family 1 tests maintained the nominal rate. Moreover, the *P* statistic had severely inflated Type I error rates when the smaller group had the larger covariances. Again, both SEM and Family 1 test statistics were able to maintain the nominal error rate in this case. Family 3 performed similarly to the standard approach in terms of both Type I error rate and power in the case of three groups. However, for two groups, Family 3 had low power, making it of questionable utility under these conditions.

With regard to power under the unequal covariance matrix conditions, SEM, followed by the Family 1 tests, had the highest values compared to the other test statistics that were able to maintain the Type I error rate at or near the nominal 0.05 level. This positive performance for SEM is in keeping with Raykov's (2001) suggestion that this approach would be particularly useful when the group covariance matrices were not equivalent. When covariance matrices were unequal, the power rates of the standard statistic, or H, could not be fairly compared because their error rates were inflated, particularly when the smaller group was paired with the covariance matrix having the larger elemental values. H had inflated error rates across most conditions. In short, when the outcome variables followed the normal distribution, SEM was able to maintain the nominal Type I error rate, and yield higher rates of power than the other methods studied here. Furthermore, in accord with Raykov (2001), the SEM approach was optimal among all the methods when the group covariance matrices were not equal and the data were normally distributed. This result supports the expectation that by allowing the group covariance matrices to be independently estimated as in SEM, it is possible to produce accurate results even when the standard assumption of homogeneity of covariance matrices is not met.

Results for procedures using trimmed estimators were similar to those that used the usual least squares estimators, with slightly lower Type I error and power

rates compared to their non-trimmed counterparts. However, these differences were consistently very small, and generally did not offer a substantive advantage over the non-trimmed test statistics. Note that power for all methods was higher in the nonnormal conditions (no differences among these three) than for normal data. At the same time, there was no concomitant inflation of the Type I error rates for a number of the test statistics when non-normal data were present. The lack of influence of non-normality may be due to the adjustments that were examined. For instance, Hotelling's T^2 is conservative with skewed distributions or when outliers are in the tails of the distribution, especially when the design is unbalanced (Everitt, 1979; Zwick, 1986). It may be that under these conditions and with adjustments such as the use of the trimmed means, the other methods remain conservative as well. Lix & Keselman (2004) state that using Family 1 with the trimmed means can result in a test that is robust to the effects of both non-normality and covariance heterogeneity. When multivariate normality is violated, the performance of Hotelling's T^2, for example, can depend on the nature of the research design and the type of departure from normality present in the data. It appears this may be the case for the other tests as well. Furthermore, other findings have suggested it may be small sample sizes with non-normal data that result in liberal results or Type I error inflation (e.g., Fouladi & Yockey, 2002; Wilcox, 1995) with these studied test statistics. Such effects with various combinations of conditions appear to deserve continued investigation to assist in sorting out when one would and would not expect a degrading of statistical power or inflation of Type I error.

Given the relative success of the Family 1 tests, it may be beneficial to take a moment and reiterate how these differ from those of the other families. Recall from the earlier discussion of this issue that Family 1 are all based on $T^2_{unequal}$, which is an analog of the univariate *t-test* calculation when group variances differ. Thus, the variances are weighted by the inverse of the group size. For tests in Families 2 and 3, the weighting of group variances was based on more complex combinations of sample size or sample proportions. Thus, the use of a simple weighting of variances by the inverse sample size may be more effective than attempting to account for the proportion of total cases in the sample, for example. Furthermore, given the very similar performances of the statistics in Family 1 to one another, it seems that the alternative methods for calculating degrees of freedom that demarcate most of these may not be particularly meaningful in conditions similar to those simulated here.

The results of this study partially supported the hypothesis that the SEM approach would have lower Type I error and higher power for all but the smallest

sample sizes. When the underlying data were normally distributed, this method would seem to be a good choice for applied researchers. SEM consistently maintained Type I error control, and yielded the highest power values, regardless of whether the group covariance matrices were equal or not. The maximum likelihood based SEM approach is closely associated with the familiar Wilks' Lambda statistic, commonly used in MANOVA testing, when the data are normally distributed, with the exception that it can be used successfully when group variances are unequal. For nonnormal data distributions simulated here, SEM was not able to maintain the nominal error rate of 0.05.

Finally, results for the trimmed methods did not differ substantially from their non-trimmed counterparts, other than by exhibiting slightly lower rejection rates. The lack of higher power in the skewed case, which was hypothesized might occur, could be due to the fact that the data were not simulated to contain true outliers, given that this was not the focus. Thus, future research should include cases where outliers are present.

Practical Recommendations for Applied Researchers

The following guideline of bullet points summarizes results; these may prove to be helpful to researchers working with MANOVA in situations where the assumptions of normality and/or equality of covariance matrices are violated. These points are organized based upon the type of assumption violation and provide the researcher with suggested test statistics to use in each situation, based upon the results of this simulation study.

1) When data are normally distributed and the groups' covariance matrices are equal, SEM provides optimal power and Type I error control.

2) When the data are not normally distributed and the groups' covariance matrices are equal, the *P* statistic maintains the nominal Type I error rate and has optimal power, whereas SEM yields an inflated Type I error, and members of Family 1 do not.

3) When the groups' covariance matrices are not equal and data are normally distributed, the *P* statistic will exhibit an inflated Type I error rate, whereas SEM, and members of the Family 1 test statistics (except for H) will maintain the nominal error rate.

4) When the groups' covariance matrices are not equal and data are normally distributed, SEM will have the highest power rates, and the Family 1 test

statistics will have lower power to find group mean differences compared to the *P*.

5) Tests based on trimmed statistics demonstrated slightly lower Type I error rates and power than their non-trimmed analogs.

Study limitations and directions for future research

As with any simulation study, there are limitations to the current work. First, a limited number of covariance inequality conditions were considered in which values for one group were multiples of those for another. Future work should expand upon the current work by investigating other covariance structures. Second, for each distribution condition, the variables had the same distribution. In practice this may not be the case, and future research should simulate situations in which variables have different distributions from one another. Third, only three non-normal distributions were considered here. Further work could, for instance, examine heavy tailed symmetric distributions, such as the Cauchy. Finally, only positively correlated dependent variables were examined here. As was noted in the introduction, the presence of negative correlations among the responses can lead to increased power for MANOVA tests. Thus, future research could extend the current work by comparing the performance of several of these methods in the presence of negative dependent variable correlations.

Conclusion

There is little doubt that with sixteen options for test statistics for MANOVA, many researchers will be overwhelmed with the choice that must be made. Many applied researchers may even be completely unaware of the various choices that exist. Furthermore, many of the choices are not available as standard options in some commonly used statistical packages, which can hinder accurate as well as wide-spread use. The result of this relative lack of access is that valid hypothesis testing in multivariate means comparisons may not be obtained when assumptions underlying the hypotheses tests are not satisfied. However, the development of the SAS macro by Lix and Keselman (2004), as well as the increasing availability of easy to use and powerful software for SEM, make many of these alternatives more accessible than ever before. Therefore, the applied researcher is encouraged to carefully consider the selection of the test given data conditions and seek resources to assist in calculations of that statistic if need be. Developers of statistical software are also encouraged to continue to integrate the various

options of these test statistics even beyond MANOVA. Though there is likely to be a lag behind development of state-of-the-art methods and software to implement these methods, researchers are encouraged to continue to attempt the use of the most appropriate method or test given the data and research question at hand. It is anticipated that the review of test statistics and results of this study will assist in guiding applied researchers in selecting optimal methods for comparing multivariate group means.

References

Algina, J., Oshima, T. C. & Tang, K. L. (1991). Robustness of Yao's, James' and Johansen's tests under variance-covariance heteroscedasticity and nonnormality. *Journal of Educational Statistics, 16,* 125-139.

Blair, R. C., Higgins, J. J., Karniski, W., & Kromrey, J. D. (1994). A study of multivariate permutation tests which may replace Hotelling's T^2 test in prescribed circumstances. *Multivariate Behavioral Research, 29,* 141-163.

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for Equality of Variances, *Journal of the American Statistical Association, 69,* 364–367.

Christensen, W. F., & Rencher, A. C., (1997). A comparison of Type I error rates and power levels for seven solutions to the multivariate Behrens–Fisher problem. *Communications in Statistics-Theory and Methods, 26,* 1251–1273.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cole, D. A., Maxwell, S. E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin, 115,* 465-474.

Coombs, W. T., Algina, J., & Oltman, D. O. (1996). Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal. *Journal of Educational and Behavioral Statistics, 66,* 137-179.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to non-normality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16-29.

Dumas, J. E., Prinz, R. J., Smith, P. E., & Laughlin, J. (1999). The EARLY ALLIANCE Prevention Trial: An Integrated Set of Interventions to Promote Competence and Reduce Risk for Conduct Disorder, Substance Abuse, and School Failure. *Clinical Child and Family Psychology Review, 2,* 37-53.

Erdfelder, E. (1981). Multivariate Rangvarianzanalyse: Ein non-parametrisches Analogonzurein- und mehrfaktoriellen MANOVA [Multivariate rank variance analysis: A nonparametric analogue for single and multivariate MANOVAs]. *TriererPsychologischeBerichte, 8.* Trier, German: Fachbereich 1-Psychologie derUniversitat Trier.

Everitt, B. S. (1979). A Monte Carlo investigation of the robustness of Hotelling's one and two sample T^2 tests. *Journal of the American Statistical Association, 74,* 48-51.

Finch, H. (2005). Comparison of the performance of the nonparametric and parametric MANOVA test statistics when assumptions are violated. *Methodology, 1,* 27-38.

Finch, W. H., & French, B. F. (2008). Testing the null hypothesis of no group mean vector difference: A comparison of MANOVA and SEM. Paper presented at the Annual meeting of the Psychometric Society, Durham, NH, June.

Fouladi, R. T., &Yockey, R. D. (2002). Type I error control of two-group multivariate tests on means under conditions of heterogeneous correlation structure and varied multivariate distributions. *Communications in Statistics – Simulation and computation, 31,* 360-378.

Hair, J. F., Anderson, R. E., Tatham, R. L, & Black, W. C. (1987). *Multivariate data analysis with readings (3rd ed).* New York, NY, Macmillan. .

Hakstian, A. R., Roed, J. C., & Lind, J. C. (1979). Two-sample T^2 procedure and the assumptionfor homogeneous covariance matrices. *Psychological Bulletin, 86,* 1255-1263.

Hancock, G. R., Lawrence, F. R., & Nevitt, J. (2001). Type I error and power of latent mean methods and MANOVA in factorially invariant and noninvariant latent variable systems. *Structural Equation Modeling, 7,* 534-556

Harris, R. J., (2001). *A primer of multivariate statistics(3rd Ed).* Mahwah, NJ: Lawrence Erlbaum.

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated non-normal distributions: Extending the Fleishman power method. *Psychometrika, 64,* 25-35.

Holloway, L. N., & Dunn, O. J. (1967). The robustness of Hotelling's T^2. *Journal of the American Statistical Association, 62,* 124-136.

Hopkins, J. W., & Clay, P. P. F. (1963). Some empirical distributions of bivariateT^2 and homoscedasticity criterion *M* under unequal variance and leptokurtosis. *Journal of the American Statistical Association, 58,* 1048-1053.

Huberty, C. L., & Morris, J. D., (1989). Multivariate analysis versus multiple univariate analysis, *Psychological Bulletin, 105,* 302-308.

Huberty, C. J, & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. New York: Wiley.

Hussein, A., & Carriere C. K. (2005) Group Sequential Procedures under Variance Heterogeneity. *Statistical Methods for Medical Research. 14*, 1-8.

Ittenbach, R. F., Chayer, D. E., Bruininks, R. H., Thurlow, M. L., & Beirne-Smith, M. (1993). Adjustment of young adults with mental retardation in community settings: comparison of parametric and nonparametric statistical techniques. *American Journal of Mental Retardation, 97,* 607-615.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratio of the population variances are unknown. *Biometrika, 41,* 19-43.

Johansen, S. (1980). The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85-92.

Johnson, R.A. &Wichern, D.W. (2002). *Applied multivariate statistical analysis.* Upper Saddle River, NJ: Prentice Hall.

Katz, B. M.. & McSweeney, M. (1980). A multivariate Kruskal-Wallis test with post hoc procedures. *Multivariate Behavioral Research, 15,* 281-297.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika, 63,* 145-163.

Kim, S. & Olejnik, S. (2004). Bias and precision of multivariate effect size measures of association for a fixed-effect analysis of variance model. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April.

Kirk, R. E. (1995). *Experimental Design: Procedures for the behavioral sciences.* New York: Wadsworth Publishing.

Korin, B. P. (1972). Some comments on the homoscedasticity criterion *M* and the multivariate analysis of variance tests T^2, *W*, and *R*. *Biometrika, 59,* 215-216.

Krishnamoorthy, K., & Xia, Y. (2006). On selecting tests for equality of two normal mean vectors. *Multivariate Behavioral Research, 41,* 533-548.

Krull, V., Choi, S., Kirk, K., Prusick, L., & French, B. F. (2010). Lexical effects on spoken-word recognition in children with normal hearing. *Ear and Hearing, 31,* 102-114.

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

Lix, L. M., & Keselman, H. J. (2004). Multivariate tests of means in independent group designs: Effects of covariance heterogeneity and nonnormality. *Evaluation in the Health Professions, 27(1),* 45-69.

McCarroll, D., Crays, N., & Dunlap, W. P. (1992). Sequential ANOVAs and type I error rates, *Educational and Psychological Measurement, 52,* 387-393.

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519-530.

Muthén, L. K., &Muthén, B. O. (1998–2004). *Mplus user's guide* (4thed.). Los Angeles: Muthén&Muthén.

Nel, D. G., & Van der Merwe, C.A., (1986). A solution to the Multivariate Behrens–Fisher problem. *Communications in Statistics-Theory and Methods, 15,* 3719–3735.

Olejnik, S. (2010). Multivariate analysis of variance. . In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods.* (pp. 328 - 328). NY: Routledge.

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

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287-312.

Pillai, K. C. S., & Jayachandran, K. (1967) Power comparisons of tests of two multivariate hypotheses based on four criteria. *Biometrika, 54,*195-210.

Puri, M. L., & Sen, P. K. (1971). *Nonparametric methods in multivariate analysis.* Malabar, FL: Krieger Publishing Company.

Ramsey, P. H. (1982). Empirical power of procedures for comparing 2 groups on p variables. *Journal of Educational Statistics, 7,* 139-156.

Raykov, T. (2001). Testing multivariable covariance structure and means hypotheses via structural equation modeling. *Structural Equation Modeling, 8(2),* 224-256.

SAS Institute. (2004). SAS software version 9.1. Cary, NC: SAS Institute.

Satorra, A., & Bentler, P.M. (1994). *Corrections to test statistics and standard enors in covariance structure analysis.* In A. von Eye & C.C. Clogg (Eds.), Latent variables analysis: Applications for developmental research (pp. 399-419). Thousand Oaks, CA: Sage.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika,40,* 87-104.

Sheehan-Holt, J. K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement, 58,* 861-881.

Stevens, J. (2001). *Applied Multivariate Statistics for the Behavioral Sciences.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

Warner**,** R. M. **(**2008**).** *Applied Statistics: From bivariate through multivariate techniques*. Thousand Oaks: Sage

Wilcox, R. R. (1995). Simulation results on solutions to the multivariate Behrens-Fisher problem via trimmed means. *The Statistician, 44,* 213-225.

Yanagihara, H., & Yuan, K-H. (2005). Three approximate solutions to the multivariate Behrens-fisher problem. *Communications in Statistics: Simulation and Computation, 34,* 957-988.

Yao, Y. (1965). An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. *Biometrika, 52,* 139-147.

Yu, C., & Muthén, B. (2002). Evaluation of model fit indices for latent variable models with categorical and continuous outcomes. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Yuen, K. K. (1974). The two-stage sample trimmed *t* for unequal population variances. *Biometrika, 61,* 165-170.

Zwick, R. (1986). Rank and normal scores alternatives to Hotelling's T^2, *Multivariate Behavioral Research, 21,* 169-186

Appendix A

The below equations supplement the material in the text so the interested reader has the formulas at their disposal. The terms are defined below and correspond to terms which appear throughout the text. For addition information on the derivation of the statistics please see the cited sources in the text.

Family 1

1) **The multivariate analog of the univariate *t*-test equation for unequal variances:**

$$T^2_{unequal} = \left(\bar{Y}_1 - \bar{Y}_2\right)' \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} \left(\bar{Y}_1 - \bar{Y}_2\right)$$

2) ***F*$_{JN}$ involves the conversion of $T^2_{unequal}$ to an *F* value:**

$$F_{JN} = \frac{T^2_{unequal}}{c_2}$$

Where

$$c_2 = p + 2C - \frac{6C}{\left(p+1\right)}$$

p = Number of outcome variables

$$C = .5\sum_{j=1}^{2} \frac{1}{n_j} tr\left(A^{-1}A_j\right)^2 + tr^2\left(A^{-1}A_j\right)$$

$$A_j = \frac{S_j}{n_j}$$

$$A = A_1 + A_2$$

This F_{JN} value for this statistic is then compared with an F critical value p,v_J degrees of freedom with $v_J = p\ (p + 2) / 3C$.

3) The F_{NV} test statistic is a transformed version of $T^2_{unequal}$:

$$F_{NV} = \frac{v_N T^2_{unequal}}{pf_2}$$

Where

$$f_2 = \left(trA^2 + tr^2 A\right)\sum_{j=1}^{2}\frac{1}{n_j - 1}\left(trA_j^2 + tr^2 A_j\right)$$

$$v_N = f_2 - p + 1$$

F_{NV} is compared to a critical F value with p, v_N degrees of freedom.

4) Yao's F_Y is based on $T^2_{unequal}$:

$$F_Y = \frac{v_K T^2_{unequal}}{pf_1}$$

Where

$$f_1 = \sum_{j=1}^{2}\frac{1}{n_j - 1}\left(\frac{T^2_{unequal}}{b_j}\right)^2$$

$$b_j = \left(\bar{Y}_1 - \bar{Y}_2\right)' V^{-1} A_j V^{-1}\left(\bar{Y}_1 - \bar{Y}_2\right)$$

$$V = A_1 + r^2 A_2 + 2rA_2^{1/2} A_2 \left(A_2^{-1/2} A_1 A_2^{-1/2}\right) A_2^{1/2}$$

$$r = \left|A_1 A_2^{-1}\right|^{1/(2p)}$$

Family 2

5) The Brown and Forsythe (F_{BF}) test statistic:

$$F_{BF} = \frac{v_{BF2}}{pf_2} T_{BF}$$

Where

$$T_{BF} = \left(\bar{Y}_1 - \bar{Y}_2\right)' \left[\left(1 - \frac{n_1}{N}\right)S_1 + \left(1 - \frac{n_2}{N}\right)S_2\right]^{-1}\left(\bar{Y}_1 - \bar{Y}_2\right)$$

$$v_{BF2} = f_2 - p + 1$$

$$f_2 = \frac{tr\left(G_1\right)^2 + tr^2\left(G_1\right)}{\frac{1}{n_{1-1}}\left[tr\left(w_1S_1\right)^2 + tr^2\left(w_1S_1\right)\right] + \frac{1}{n_{2-1}}\left[tr\left(w_2S_2\right)^2 + tr^2\left(w_2S_2\right)\right]}$$

$$w_j = 1 - \frac{n_j}{N}$$

$$G_1 = w_1S_1 + w_2S_2$$

$$v_{BF1} = \frac{tr\left(G_1\right)^2 + tr^2\left(G_1\right)}{tr\left(G_2\right)^2 + tr^2\left(G_2\right) + tr\left(\sqrt{w_1S_1}\right)^2 + tr\left(\sqrt{w_2S_2}\right)^2 + tr^2\left(\sqrt{w_1S_1}\right) + tr^2\left(\sqrt{w_2S_2}\right)}$$

$$G_2 = \frac{n_1}{N}S_1 + \frac{n_2}{N}S_2$$

6) The Kim (*FK*) test statistic:

$$F_K = \frac{v_k\left(\bar{Y}_1 - \bar{Y}_2\right)' V^{-1}\left(\bar{Y}_1 - \bar{Y}_2\right)}{c_1 m f_1}$$

Where

$$c_1 = \frac{\sum_{j=1}^{2} h_1^2}{\sum_{j=1}^{2} h_1}$$

$$h_1 = \frac{\left(d_1 + 1\right)}{\left(d_1^{1/2} + r\right)^2}$$

$$m = \frac{\left(\sum_{j=1}^{2} h_1\right)^2}{\sum_{j=1}^{2} h_1^2}$$

$$v_k = f_1 - p + 1$$

7) Winsorized variance:

$$S_{Wp}^2 = \frac{\sum_{i=1}^{n}\left(Z_i - \bar{Y}_{wp}\right)^2}{n-1}$$

Where

$\bar{Y}_{wp} =$ Winsorized mean of variable *p*

$Z_i = Y_{L+1}$ if $Y_i \le Y_L$

$Z_i = Y_{H-1}$ if $Y_i \geq Y_H$

Otherwise $Z_i = Y_i$

Y_L = Lower cut score corresponding to 20th percentile value.

Y_H = Upper cut score corresponding to 80th percentile value.

8) T^2 and $T^2_{unequal}$ **can be calculated using the trimmed means and Winsorized covariance matrices as:**

$$T_R^2 = \left(\bar{Y}_{T1} - \bar{Y}_{T2}\right)' \left[S_w \left(\frac{1}{h_1} + \frac{1}{h_2} \right) \right]^{-1} \left(\bar{Y}_{T1} - \bar{Y}_{T2}\right)$$

Where

$$S_w = \frac{(n_1 - 1)}{(h_1 - 1)} S_{w1} + \frac{(n_2 - 1)}{(h_2 - 1)} S_{w2}$$

$\bar{Y}_{Tj}$ = Trimmed mean for group *j*

h_j = Number of group *j* that is kept after trimming.

9) A version of Hotelling's T^2 that does not use the pooled covariance matrix:

$$T^2_{R\ unequal} = \left(\bar{Y}_{T1} - \bar{Y}_{T2}\right)' \left(\frac{(n_1 - 1)}{(h_1 - 1)h_1} S_{w1} + \frac{(n_2 - 1)}{(h_2 - 1)h_2} S_{w2} \right)^{-1} \left(\bar{Y}_{T1} - \bar{Y}_{T2}\right)$$

Family 3

10) Rank based nonparametric test

Convert Pillai's trace value using ranks into the chi-square statistic: $\chi^2 = (n-1)P$ where *P* is Pillai's trace and *n* is the total sample size. Compare the value with the χ^2 distribution with *k* (*p* – 1) degrees of freedom, where *k* is the number of groups for the independent variable and *p* is the number of response variables.

Family 4

11) Structural Equation Model based test

To test of the null hypothesis of group differences on the responses is the difference in the χ^2 fit statistics: $\chi^2_{\text{Constrained}} - \chi^2_{\text{Unconstrained}}$. Allowing the group means to differ results in a saturated CFA model so that the value of $\chi^2_{\text{Unconstrained}} = 0$.

The test of the null hypothesis of group differences across the vector of dependent variable means is equivalent to $\chi^2_{\text{Constrained}} - \chi^2_{\text{Unconstrained}} = \chi^2_{\text{Constrained}} - 0 = \chi^2_{\text{Constrained}}$.