A MONTE CARLO STUDY OF SEVERAL MULTIVARIATE ANALYSIS OF VARIANCE PROCEDURES

By

CARLA ANNE REICHARD

Bachelor of Science With High Honors
University of Oklahoma
1980

Master of Arts
University of Kansas
1982

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
In partial fulfillment of
The requirements for
The Degree of
DOCTOR OF PHILOSOPHY
May, 1999

A MONTE CARLO STUDY OF SEVERAL

MULTIVARIATE ANALYSIS OF

VARIANCE PROCEDURES

Thesis Approved:

William Thomas Coombs

Thesis Adviser

Katye M. Lesser

Laura L. B. Barnes

Donald Bear

Wayne B. Powell

Dean of the Graduate College

ACKNOWLEDGMENTS

It has been six long years since I began my doctoral program, and in that time a number of things have changed: a new family member, a new house, and a new job. Through all of these changes, several things have kept me going when things got tough.

One is the quality of the doctoral program itself. To my committee members, Katye Perry, Tom Coombs, Laura Barnes, and Ron Beer, and to the other members of the research and higher education programs: thank you for adhering to the highest standards of professionalism, ethics, and pedagogy, and for making every class full of useful knowledge. Special thanks are due to Tom for having the patience to guide me through the long, steep climb involved in learning the details of Monte Carlo studies. It was worth the trouble, at least from my end.

Another is the flexibility and commitment to professional development shown by the Office of Planning, Budget and Institutional Research, and especially my boss, Joe Weaver. I hope that I have shown, and will continue to show, that this commitment will benefit the office and OSU as much as it has personally benefited me.

Finally, I would like to thank my husband, Kouider, and sons Adam and Ben, for sacrifices above and beyond the call of duty, and my parents, Jim and Judy Reichard, for many hours of babysitting at crucial times through the years.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

α	nominal Type I error rate
t_v	Welch v statistic
s_p^2	pooled sample variance
σ_i	population variance
s_i	sample variance
G	number of treatment groups
P	number of dependent variables
T^2	Hotelling's test statistic
S	sample covariance matrix
H	pooled between group sum of squares and cross products matrix
E	pooled error sum of squares and cross products matrix
R	Roy's largest root criterion
U	Hotelling-Lawley trace criterion
L	Wilk's likelihood ratio criterion
V	Pillai-Bartlett trace criterion
J	James test statistic
R*	Coombs-Algina R* statistic
U1*	test statistic for Coombs-Algina U1* test
U2*	test statistic for Coombs-Algina U2* test

L*	test statistic for Coombs-Algina L* test
V*	test statistic for Coombs-Algina V* test
Rho	degree of dependence within subgroups
Nspsub	number of subjects per subgroup
Nsubpgr	number of subjects per treatment group
NCP	type of noncentrality
Effect Size	measure of the degree of difference between groups

Chapter 1

Introduction

Many of the most common inferential statistical techniques in use today fit under the umbrella designation of the Generalized Linear Model (GLM) (Agresti, 1996). What do these techniques have in common? All assume that a set of dependent variables ($Y_1, Y_2, \ldots, Y_n$) can be modeled by a set of linear equations in *m* independent variables, as follows:

$$Y_1 = a_1 + b_{11}X_{11} + b_{21}X_{21} + \ldots + b_{m1}X_{m1} + \varepsilon_1$$

$$Y_2 = a_2 + b_{12}X_{12} + b_{22}X_{22} + \ldots + b_{m2}X_{m2} + \varepsilon_2$$

.

.

.

$$Y_n = a_n + b_{1n}X_{1n} + b_{2n}X_{2n} + \ldots + b_{mn}X_{mn} + \varepsilon_n$$

When the dependent variables are assumed to be continuous, the model is referred to as the General Linear Model. Techniques in this model include the independent samples *t* test, analysis of variance (ANOVA), analysis of covariance (ANCOVA), multiple regression, multivariate analysis of variance (MANOVA), and multivariate analysis of covariance (MANCOVA). Though some of these techniques (and their variations, such as repeated measures) require specialized assumptions, all contain the following three assumptions:

1) Each Y_i (*i=1,...,n*) is sampled from a normal population.

2) The *n* population variances are equal.

3) Each observation Y_{ij} ($i=1,...,n; j=1,...,N$) is independent of all other observations.

The Problem

In the field of applied statistics, much research has been devoted to the topic of what researchers should do when, as often happens in the real world, experimental data does not completely conform to the assumptions required by the model. Glass, Peckham, and Sanders (1972) observe that little of what we know about what happens to inferential statistical procedures under non-ideal conditions is due to mathematical proofs. Instead, much useful information comes from Monte Carlo studies of various conditions which violate the assumptions of the statistical model. Modern computers have increased the use of this technique, and today, there is a substantial body of literature detailing the consequences of the violation of assumptions for many of the tests falling under the General Linear Model.

In addition to examining what happens when the given assumptions are violated for a particular test, researchers have pursued the development of alternatives to statistical tests within the General Linear Model. In the 1950s and 1960s, non-parametric alternatives to common tests were developed which did not require the assumption that the sampled populations be normally distributed (Conover, 1981). More recently, development has centered on parametric alternatives which do not require the assumption of homogeneity of variance. These alternatives may eventually replace the most common techniques taught

today. Before this can happen, researchers must build up a substantial body of knowledge comparing the different tests.

Robustness and Power: Type I and Type II Error

According to Keselman, Lix, and Keselman (1996), Box (1953) coined the term robustness to refer to insensitivity of a statistical test's rates of Type I error and power to violations of its derivational assumptions. In robustness studies, a "nominal" Type I error rate is set (e.g., at $\alpha = .05$), and a large number of trials are run, sampling from the same specified population (for example, a skewed, rather than a normal, population). The estimated "actual" Type I error rate (the proportion of trials in which the test detects a false difference) is then compared to the nominal rate. ANOVA, for example, is said to be robust to violations of the assumption of normality, because a number of Monte Carlo studies have found that the actual Type I error rate does not vary a tremendous amount from the nominal Type I error rate. This is good news, for it means that researchers using data from non-normal populations can safely interpret significant results from ANOVA.

Power, on the other hand, refers to the ability of the statistical test to detect a true difference (the avoidance of Type II error). For any given statistical test and a fixed sample size, Type I and Type II error rates are inversely related; that is, decreasing the Type I error rate (i.e., setting a very stringent alpha level) will decrease the probability of detecting a true difference (Type II error), and thus will decrease power. The researcher who has settled on a given technique must thus strike a balance between Type I and Type II error rates.

The developer of new statistical techniques, on the other hand, wants to find new tests which will be both robust and powerful when compared to existing tests. Certainly, a test which eliminates one assumption appears to have a good head start on the process. However, the group of non-parametric tests, which eliminate the assumption of normality, have lost popularity since Monte Carlo studies have shown that (1) they are largely unnecessary, since in most cases, parametric tests are robust to violations of normality; and (2) they have lower power than the non-parametric tests (Glass, Peckham, & Sanders, 1972). As we shall see, the parametric tests are more sensitive to violations of the assumption of homogeneity of variance, prompting the development of another group of alternative tests.

Parametric Alternatives to Tests Within the General Linear Model

The "Behrens-Fisher problem" is the name generally used to refer to heteroscedasticity in problems using the statistical tests contained in the general linear model. Heteroscedasticity refers to the condition in which different experimental groups are sampled from populations with unequal variances. The term "Behrens-Fisher" comes from the work of Behrens (1929), who found a solution to the problem of heteroscedasticity; and Fisher (1935), who showed that Behrens's solution could be derived from Fisher's fiducial principle. Today there are a number of solutions to the Behrens-Fisher problem. In general, these alternative tests compare favorably with their counterparts in the General Linear Model. However, none are commonly used. Only a few (such as the Welch test) are available on the most widely used statistical software packages

such as SAS and SPSS, and they are often omitted in beginning statistical textbooks. It is not clear why this is the case, though perhaps it is believed that their advantages, though well-documented, are not large enough to warrant a wholesale change. In the case of parametric alternatives to MANOVA, there simply are not enough Monte Carlo studies of alternatives completed yet to have a clear picture of which alternative tests, if any, are superior.

Purpose of the Study

The purpose of this study is to compare the Type I error rates and powers of selected alternatives to MANOVA under a variety of conditions. The Pillai-Bartlett (Bartlett, 1939; Pillai, 1955) test, the Johansen (1980) test, and four Coombs-Algina (1996) tests will be compared under conditions of non-normality, heteroscedasticity, and dependence of observations. The specific conditions have been chosen with two basic criteria in mind: (1) to mimic conditions which violate the stated assumptions of the General Linear Model, but which might reasonably be expected to occur with some frequency in educational research; and (2) to cover a range of conditions, such as the number of dependent variables, and the number of groups, that will either lend confidence that our conclusions are stable, or will point out that the choice of a "best" alternative test is dependent on the conditions themselves.

Two questions have emerged to guide the research in this study.

Research Question 1. Do rejection rates differ as a function of the statistical test used, total sample size, the number of groups, the number of

dependent variables, the number of subjects per subgroup, the intraclass correlation, the population effect size, and the type of noncentrality?

Research Question 2. Under what conditions does each test maintain adequate control of Type I error rate and have suitable power?

Significance of the Study

Educational researchers who choose quantitative methods, and who believe that their research will lead to improved teaching and learning, are dependent on both the robustness and the power of the statistical tests they choose. Since the researcher searching for an inferential test is only sampling from a larger population, (s)he usually does not know the true population parameters. A non-robust test, applied to samples from populations which violate the chosen test's assumptions, may lead the researcher to conclude that there is a difference between groups when none exists. A test with low power may fail to detect a difference when one does exist. In either case, a substantial amount of effort is wasted.

Stevens (1992) notes that research designs using multiple dependent variables are common in education, because educational treatments often affect more than one variable at the same time, and also points out that in many situations, the use of a multivariate test is preferable to the use of separate univariate analyses when multiple dependent variables are being measured. Thus, there is a need for robust, powerful tests which will detect the presence of group differences on multiple dependent variables. As we shall see, MANOVA, like ANOVA, suffers from non-robustness under certain conditions, most notably

heterogeneity of variance. It is therefore important to know whether any of the several MANOVA alternatives which have been developed will perform better than MANOVA under violations of assumptions that are likely to be encountered in research practice. This study will extend the knowledge of the performance of several of these tests under a variety of conditions.

Chapter 2

Review of Literature

Historically, development of statistical tests has followed a pattern beginning with examination of differences between two groups, then moving on to differences between several groups; and with examination of group differences using one dependent variable, then moving on to examination of group differences using several dependent variables. The recognition that each test is a special case of a more general pattern, in this case what we have referred to as the General Linear Model, came later still. The development of tests which are alternatives to the General Linear Model followed a similar pattern. This chapter traces the development of both the General Linear Model tests and their parametric alternatives in that same order. After all of the tests are presented, the literature which deals with their comparative effectiveness is examined. It will be seen that many of the original ideas for alternatives to the simple *t* test have been extended to accommodate several groups and to accommodate several dependent variables. Some generalizations may also be drawn about the relative robustness of each type of solution.

Univariate Tests Comparing Two Means

The independent samples *t* test is used to compare two means when independent random samples have been drawn from two normal populations with equal variances. The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

where the pooled variance is computed as

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Welch (1938) developed several alternatives to the independent samples *t* test which do not require the assumption of equal variances for both groups. The Welch *v* statistic, or t_v, is calculated as

$$t_v = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The critical value must be determined by the use of either: a) approximate degrees of freedom (APDF) solutions or b) series solutions. The APDF solutions approximate the degrees of freedom which define the sampling distribution. Series solutions are derived by utilizing a series expansion to determine the critical value to be used. The degrees of freedom for the APDF solutions are given by

$$f = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{\sigma_1^4}{n_1^2(n_1 - 1)} + \frac{\sigma_2^4}{n_2^2(n_2 - 1)}}.$$

In practice, σ_i is estimated by s_i, the sample variance. The first three terms of critical values for the series solutions are shown in Table 1. The zero-order test is often called the asymptotic test. The first- and second-order tests are generally referred to as James's first-order and second-order test, since James (1951, 1954) generalized those terms of Welch's series solutions to both the G-sample case (an alternative to ANOVA) and the multivariate case (an alternative

Table 1. The First Three Terms of the Critical Value for t_ν

Power of $(n_i - 1)^{-1}$	Term
Zero	z
One	$z\left[\dfrac{(1+z^2)\displaystyle\sum_{i=1}^{2}\dfrac{\left(\frac{s_i^2}{n_i}\right)^2}{n_i-1}}{\left[\displaystyle\sum_{i=1}^{2}\dfrac{s_i^2}{n_i}\right]^2}\right]$
Two	$z\left[-\dfrac{1+z^2}{2}\dfrac{\displaystyle\sum_{i=1}^{2}\left(\dfrac{s_i^2}{n_i(n_i-1)}\right)^2}{\left[\displaystyle\sum_{i=1}^{2}\dfrac{s_i^2}{n_i}\right]^2}\right.$
	$\left.+\dfrac{3+5z^2+z^4}{3}\dfrac{\displaystyle\sum_{i=1}^{2}\left[\dfrac{\left(\frac{s_i^2}{n_i}\right)^3}{n_i-1}\right]^2}{\left[\displaystyle\sum_{i=1}^{2}\dfrac{s_i^2}{n_i}\right]^3}\right.$
	$\left.-\dfrac{15+32z^2+9z^4}{32}\dfrac{\displaystyle\sum_{i=1}^{2}\left[\dfrac{\left(\frac{s_i^2}{n_i}\right)^2}{n_i-1}\right]^2}{\left[\displaystyle\sum_{i=1}^{2}\dfrac{s_i^2}{n_i}\right]^4}\right]$

to MANOVA). Aspin (1948) investigated the third- and fourth-order terms. The APDF solution is commonly known simply as the Welch test.

Yeun (1974) developed a variation of the Welch test based on trimmed means and Windsorized variances, to improve power when sampling from long-tailed symmetric distributions. The *j*-times trimmed mean is defined by omitting the *j* highest and *j* lowest scores, and dividing by *(n-2j)*, rather than *n*. The j-times Windsorized mean is obtained by replacing the scores omitted from the trimmed mean with *(j+1)* times the lowest and highest scores remaining:

$$\bar{x}_{wj} = \frac{1}{n}\{(j+1)x_{j+1} + x_{j+2} + \ldots + x_{n-j-1} + (j+1)x_{n-j}\}.$$

The j-times Windsorized variance is defined in a similar manner:

$$s^2_{wj} = \frac{(j+1)\left(x_{j+1} - \bar{x}_{wj}\right)^2 + \left(x_{j+2} - \bar{x}_{wj}\right)^2 + \ldots + \left(x_{n-j-1} - \bar{x}_{wj}\right)^2 + (j+1)\left(x_{n-j} - \bar{x}_{wj}\right)^2}{n-2j}.$$

Yeun's test statistic is

$$t^*_v = \frac{\bar{x}_{tj1} - \bar{x}_{tj2}}{\sqrt{\frac{s^2_{wj1}}{(n_1 - 2j_1)} + \frac{s^2_{wj2}}{(n_2 - 2j_2)}}}.$$

The critical value is a percentile of Student's *t* distribution with f_t degrees of freedom, where

$$f_t = \frac{\left[\frac{s^2_{wj1}}{h_1} + \frac{s^2_{wj2}}{h_2}\right]^2}{\frac{\left[\frac{S^2_{wj1}}{h_1}\right]^2}{h_1 - 1} + \frac{\left[\frac{s^2_{wj2}}{h_2}\right]^2}{h_2 - 1}},$$

and $h_i = n_i - 2j_i$.

Wilcox (1992) defined the *H* test based on the one-step *m*-estimator of

location:

$$H = \frac{\bar{x}_{m1} - \bar{x}_{m2}}{\sqrt{s_{m1}^2 + s_{m2}^2}},$$

where $\bar{x}_{mi}$ is the one-step *m*-estimator in the *i*th group and s_{mi}^2 is the estimated sampling variance of $\bar{x}_{mi}$. To define $\bar{x}_m$, let *MAD* be the median absolute deviation, and let $\zeta = MAD \div .6475$, let j_1 be the number of observations for which $(x - M)/\zeta < -1.28$, and j_2 be the number of observations for which $(x - M)/\zeta > 1.28$. Then the one-step *m*-estimator used by Wilcox is

$$\bar{x}_m = \frac{1.28\zeta(j_2 - j_1) + x_{j_1+1} + \ldots + x_{n-j_2}}{n - j_1 - j_2}.$$

Let $\Psi(x) = \max\{-k, \min(k, x)\}$. Then

$$s_m^2 = \frac{\zeta^2 \Sigma \Psi^2[(x - M)/\zeta]}{\{\Sigma \Psi'[(x - M)/\zeta]\}^2}.$$

Wilcox employed a bootstrap procedure to calculate the critical value for *H*; the reader is referred to Wilcox (1992) for details.

Univariate Tests Comparing *G* Means

The ANOVA *F* test compares *G* means ($G \geq 2$) when independent random samples have been drawn from normal populations with equal variances. The test statistic is

$$F = \frac{MS_b}{MS_w},$$

where

$$MS_b = \frac{\sum_{i=1}^{G} n_i (\bar{x}_i - \bar{x}_.)^2}{G - 1},$$

$$MS_w = \frac{\sum_{i=1}^{G}(n_i - 1)s_i^2}{N - G},$$

and

$$\bar{x}_{\cdot} = \frac{1}{G}\sum_{i=1}^{G}\bar{x}_i .$$

The ANOVA test statistic has an *F* distribution with *G-1* and *N-G* degrees of freedom, where *N* is the size of total sample.

Welch (1951) generalized the Welch (1947) APDF solution for *G* groups as follows:

$$F_v = \frac{\sum_{i=1}^{G} w_i(\bar{x}_i - \bar{x})^2 / (G-1)}{1 + \frac{2(G-2)}{G^2 - 1}\sum_{i=1}^{G}\frac{1}{n_i - 1}\left(1 - \frac{w_i}{w}\right)^2},$$

where

$$w_i = \left[\frac{s_i^2}{n_i}\right]^{-1}, \; i=1,\ldots,G,$$

$$w = \sum_{i=1}^{G} w_i \text{, and}$$

$$\bar{x} = \sum_{i=1}^{G}\frac{w_i \bar{x}_i}{w}.$$

F_v is approximately distributed as *F* with *G*-1 and f_1 degrees of freedom, where

$$f_1 = \left[\frac{3}{G^2 - 1}\sum_{i=1}^{G}\frac{i}{n_i - 1}\left(1 - \frac{w_i}{w}\right)^2\right]^{-1}.$$

James (1951) generalized the Welch (1947) series solutions using $J = \sum_{i=1}^{G} w_i(\bar{x}_i - \bar{x})^2$, where w_i, $\bar{x}_i$, and $\bar{x}$ are defined as above in the Welch APDF solution. For the zero-order or asymptotic test, the critical value is a percentile of

a chi-square distribution with *G*-1 degrees of freedom. This distribution will not accurately approximate the sampling distribution of the test statistic unless sample sizes are sufficiently large. The James first- and second-order tests adjust for this problem; the second-order test is computationally intensive and is not presented here. The reader is referred to James (1951) for details. Oshima & Algina (1992a) wrote a computer program which computes the second-order test. The first order test computes

$$2h\left(s_i^2\right) = \chi^2_{G-1;\alpha}\left[1+\frac{3\chi^2_{G-1;\alpha}+G+1}{2\left(G^2-1\right)}\sum_{i=1}^{G}\frac{1}{n_i-1}\left(1-\frac{w_i}{w}\right)^2\right];$$

if $J \geq 2h\left(s_i^2\right)$ then the null hypothesis is rejected.

Brown and Forsythe (1974) extended the Welch (1947) APDF test to *G* groups somewhat differently than Welch (1951), and proposed the test statistic

$$F^* = \frac{\sum_{i=1}^{G} n_i\left(\bar{x}_i - \bar{x}_.\right)^2}{\sum_{i=1}^{G}\left(1-\frac{n_i}{N}\right)s_i^2},$$

approximately distributed as *F* with *G*-1 and g_2 degrees of freedom, where

$$g_2 = \frac{\left[\sum_{i=1}^{G}\left(1-\frac{n_i}{N}\right)s_i^2\right]^2}{\sum_{i=1}^{G}\frac{\left[\left(1-\frac{n_i}{N}\right)s_i^2\right]^2}{n_i-1}}.$$

For two groups, the Brown-Forsythe and the Welch (1951) APDF tests are both equivalent to the Welch (1947) APDF test.

Rubin (1982) extended the Brown-Forsythe test by replacing the first degree of freedom, *G*-1, with $g_1 = \left[\sum_{i=1}^{G}\left(1-\frac{n_i}{N}\right)s_i^2\right]^2\left[\left[\sum_{i=1}^{G}\frac{n_i}{N}s_i^2\right]^2 + \sum_{i=1}^{G}\left(1-2\frac{n_i}{N}\right)s_i^2\right]^{-1}$.

Wilcox has developed two alternatives to ANOVA which are generalizations of tests from the two-group case. The first (Wilcox, 1995) is a generalization of Yeun's test:

$$F_v^* = \frac{\sum_{i=1}^{G} w_i\left(\bar{x}_{tji} - \bar{x}_t\right)^2 / (G-1)}{1+\frac{2(G-2)}{G^2-1}\sum_{i=1}^{G}\frac{1}{h_i-1}\left(1-\frac{w_i^*}{w^*}\right)^2},$$

where

$$w_i^* = \frac{h_i^2(h_i-1)}{(n_i-1)s_{wji}^2},$$

$$w^* = \sum_{i=1}^{G} w_i^* \text{, and}$$

$$\bar{x}_t = \frac{1}{w^*}\sum_{i=1}^{G} w_i^*\bar{x}_{tji}.$$

The statistic F_v^* is approximately distributed as *F* with *G*-1 and f_t^* degrees of freedom, where

$$f_t^* = \left[\frac{3}{G^2-1}\sum_{i=1}^{G}\frac{1}{h_i-1}\left(1-\frac{w_i^*}{w^*}\right)^2\right]^{-1}.$$

Wilcox (1993) also generalized his H test to G groups, using

$$Z = \frac{1}{N}\sum_{i=1}^{G} n_i\left(\bar{x}_{im} - \bar{x}_m\right),$$

where

$$\bar{x}_m = \sum_{i=1}^{G}\bar{x}_{im} / G.$$

The critical value of Z is determined by bootstrap methods; the reader is referred to Wilcox (1993) for details.

Alexander and Govern (1994) developed the test statistic

$$A = \sum_{i=1}^{G} z_i^2 ,$$

where

$$z_i = c + \frac{c^3 + 3c}{b} - \frac{4c^7 + 33c^5 + 240c^3 + 855c}{10b^2 + 8bc^4 + 1000b},$$

$$a = n_i - 1.5 ,$$

$$b = 48a^2 ,$$

$$c = \left[a \ln\left(1 = \frac{t_i^2}{n_i - 1} \right) \right]^{\frac{1}{2}},$$

$$t_i = \frac{\bar{y}_i - y^+}{s_i},$$

$$y^+ = \sum_{i=1}^{G} w_i \bar{y}_i ,$$

$$w_i = \frac{\frac{1}{s_i^2}}{\sum_{i=1}^{G} \frac{1}{s_i^2}},$$

and

$$s_i = \left[\frac{\sum_{j=1}^{n_i} \left(y_j - \bar{y}_i \right)^2}{n_i \left(n_i - 1 \right)} \right]^{\frac{1}{2}} .$$

The statistic A is approximately distributed as χ^2 with $G-1$ degrees of freedom.

Multivariate Tests Comparing Two Means

Hotelling's (1931) T^2 compares two mean vectors when independent random samples have been drawn from multivariate normal populations with equal variance-covariance matrices. The test statistic is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right)' \mathbf{S}^{-1}\left(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2\right),$$

where

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2},$$

$\bar{\mathbf{x}}_i$ is the sample mean vector for the *i*th group, and $\mathbf{S}_i$ is the sample covariance matrix for the *i*th group. Then if *p* is the number of dependent variables,

$$\frac{n_1 + n_2 - p - 1}{(n_1 + n_2 - 2)p} T^2$$

has an *F* distribution with *p* and $n_1 + n_2 - p - 1$ degrees of freedom.

There are several alternatives to T^2 which do not require the assumption of equal covariance matrices. The James (1954) first- and second-order tests, Johansen (1980) test, Nel and van der Merwe's (1986) test and Yao's (1965) test are all based on the following test statistic:

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

They differ only in their critical values. James (1954) generalized the James first- and second-order tests which are alternatives to the independent samples *t* test. The first-order test uses critical value $\chi^2_{\alpha;p}\left(B + \chi^2_{\alpha;p} C\right)$ where

$$B = 1 + \frac{1}{2p}\sum_{i=1}^{G}\frac{tr^2\left(\mathbf{V}^{-1}\mathbf{A}_i\right)}{n_i - 1}$$

and

$$C = \frac{1}{p(p+2)}\left[\sum_{i=1}^{G}\frac{tr\left(\mathbf{V}^{-1}\mathbf{A}_i\right)^2}{n_i - 1} + \frac{1}{2}\sum_{i=1}^{G}\frac{tr^2\left(\mathbf{V}^{-1}\mathbf{A}_i\right)}{n_i - 1}\right],$$

where *tr* is the trace operator, $\mathbf{A}_i = \mathbf{S}_i / n_i$, and $\mathbf{V} = \mathbf{A}_1 + \mathbf{A}_2$. James's second-order test is computationally intense and is not presented here; the reader is referred to James (1954) for details.

Johansen's (1980) test is a generalization of the Welch (1947) APDF test with test statistic T_v^2 / c_1 where

$$c_1 = p + 2A - \frac{6A}{p+2}$$

and

$$A = \sum_{i=1}^{2}\frac{tr\left(\mathbf{V}^{-1}\mathbf{A}_i\right)^2 tr^2\left(\mathbf{V}^{-1}\mathbf{A}_i\right)}{2(n_i - 1)}.$$

The critical value is a percentile of the *F* distribution with *p* and $p(p+2)/3A$ degrees of freedom.

Another generalization of the Welch APDF test was developed by Yao (1965). The test statistic is

$$F_v = \frac{f_2 - p + 1}{pf_2}T_v^2,$$

where

$$\frac{1}{f_2} = \sum_{i=1}^{2}\frac{1}{(n_i - 1)}\left(\frac{v_i}{T_v^2}\right)^2,$$

$$v_i = \bar{\mathbf{y}}'\mathbf{V}^{-1}\mathbf{A}_i\mathbf{V}^{-1}\bar{\mathbf{y}},$$

and

$$\bar{\mathbf{y}} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 .$$

Nel and van der Merwe (1986) presented yet another generalization of the Welch APDF solution, using the test statistic

$$F_v = \frac{f_3 - p + 1}{p f_3} T_v^2 ,$$

where

$$f_3 = \frac{tr\mathbf{V}^2 + tr^2\mathbf{V}}{\sum_{i=1}^{2} \frac{tr\mathbf{A}_i^2 + tr^2\mathbf{A}_i}{n_i - 1}} .$$

The critical value is a percentile of the *F* distribution with *p* and $f_3 - p + 1$ degrees of freedom.

A different alternative to T^2 was presented by Kim (1992). The test statistic is

$$G = \frac{f_2 - p + 1}{c_2 m f_2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \mathbf{A}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) ,$$

where

$$\mathbf{A} = \mathbf{A}_1 + r^2 \mathbf{A}_2 + 2r \mathbf{A}_2^{1/2} \left(\mathbf{A}_2^{-1/2} \mathbf{A}_1 \mathbf{A}_2^{-1/2} \right)^{1/2} \mathbf{A}_2^{1/2} ,$$

$$r = \left| \mathbf{A}_1 \mathbf{A}_2^{-1} \right|^{1/2p} ,$$

$$c_2 = \sum_{j=1}^{p} L_j^2 \Big/ \sum_{j=1}^{p} L_j ,$$

$$m = \left(\sum_{j=1}^{p} L_j \right)^2 \Big/ \sum_{j=1}^{p} L_j^2 ,$$

$$L_j = \left(d_j + 1\right) / \left(d_j^{1/2} + r\right)^2,$$

d_j is the jth eigenvalue of $\mathbf{A}_1\mathbf{A}_2^{-1}$, and f_2 is calculated as in Yao's procedure. The critical value is a fractile of the *F* distribution with *m* and $f_2 - p + 1$ degrees of freedom.

Comparisons of Performance for the ANOVA *F* Test and Alternatives

The ANOVA *F* test is not robust with respect to violations of the assumption of homogeneity of variance (Brown & Forsythe, 1974; Clinch & Keselman, 1982; Harwell, Rubinstein, Hayes, & Olds, 1992; Kohr & Games, 1974; Rogan & Keselman, 1977; Tomarken & Serlin, 1986; Wilcox, 1988; Wilcox, Charlin & Thompson, 1986). In other respects, however, its performance generally parallels that of the *t* test. It is known to be conservative under the positive condition, and liberal under the negative condition (Box, 1954a; Brown & Forsythe, 1974; Clinch & Keselman, 1982; Horsnell, 1953; Rogan & Keselman, 1986; Wilcox, 1988; Wilcox, Charlin & Thompson, 1986). When comparing the Type I error rates of *F* with the alternative Brown-Forsythe test, the James first- and second-order tests, the Rubin test, the Welch test, Alexander and Govern's test, and the Wilcox *Z*, when sampling from normal populations with heterogeneity of variance, the following observations can be made:

(a) Each alternative is superior to *F* (Brown & Forsythe, 1974; Clinch & Keselman, 1982; Rubin, 1982; Wilcox, 1988; Wilcox, Charlin & Thompson, 1986).

(b) The Welch test and the Brown-Forsythe test are generally competitive with one another, and both outperform the James first-order test (Brown & Forsythe, 1974).

(c) Rubin's test is comparable to the Welch test and performs better than the Brown-Forsythe test (Rubin, 1982).

(d) The James second-order test outperforms both the Brown-Forsythe and Welch tests under the greatest variety of conditions (Dijkstra & Werter, 1981; Wilcox, 1988).

(e) The above suggests that the James second-order test will outperform Rubin's test.

(f) The Alexander-Govern test is comparable to the James second-order test (Alexander & Govern, 1994).

Wilcox (1988) suggests another advantage of the James second-order test: the Welch test can be liberal even when the assumption of homogeneity of variance is met; and the Brown-Forsythe and Welch tests can both be liberal when sample sizes are equal but variances are not equal. The James second-order test did not exhibit these problems.

When the condition of normality is violated, actual Type I error rates can be negatively affected for the Welch test, the Brown-Forsythe test, and the James first- and second-order tests (Oshima & Algina, 1992b). Skewed distributions have a particularly strong affect on the Welch test and the James second-order test. For symmetric distributions, we see results similar to those found for the *t* test and its alternatives: actual Type I error rates can be too high

for short-tailed distributions, and too low for long-tailed distributions. For the Brown-Forsythe test, the effect of distribution on actual Type I error rates is more variable. The Wilcox F_v^* test, which is a generalization of Yuen's (1974) trimmed-means *t*, decreases Type I error rates relative to the Welch test, the Brown-Forsythe test, and the James first- and second-order tests under nonnormal conditions (Wilcox, 1993b). Wilcox's *Z* test has actual Type I error rates close to nominal Type I error rates for nonnormal as well as normal distributions (Wilcox, 1993a). Alexander and Govern (1994) did not study conditions of nonnormality for their test.

The ANOVA *F* test is uniformly the most powerful test when all assumptions are met; however, the Brown-Forsythe, Welch, James first-order and James second-order tests all have power close to that of *F* under those conditions (Brown & Forsythe, 1974; Dijkstra & Werter, 1981; Tomarken & Serlin, 1986; Wilcox, Charlin & Thompson, 1986). For three or four groups, Tomarken & Serlin (1986) studied how the Brown-Forsythe and Welch tests compare with different groupings of population means and variances. They found that the Brown-Forsythe test was more powerful when there was one extreme population mean paired with a large variance. The Welch test was more powerful when (a) the population means were equally spaced, (b) there was one large population mean, one small population mean, and two equal intermediate means, and (c) there was one extreme population mean paired with a small variance. Dijkstra & Werter (1981) found that the comparison of power between the James second-

order test and the Brown-Forsythe test was similarly affected by the pairing of different population means and variances.

Comparisons of Performance for Hotelling's T^2 Test and Alternatives

Like the ANOVA *F* test, the performance of Hotelling's T^2 under conditions of heterogeneity of the covariance matrices depends on sample size and whether or not sample sizes are equal. If $n_1 = n_2$ and both are sufficiently large, and if populations are multivariate normal, then T^2 is relatively robust to violations of the homogeneity assumption (Holloway & Dunn, 1967; Hopkins & Clay, 1963; Ito & Schull, 1964). When sample sizes are not equal, heteroscedasticity can strongly affect the actual Type I error rate even when the degree of sample size inequality is small (Algina & Oshima, 1990). When the larger samples are selected from populations with greater dispersion (the positive condition), T^2 is conservative (Holloway & Dunn, 1967; Hopkins & Clay, 1963; Hakstian, Roed, & Lind, 1979) and power is lower than it would be under conditions of homoscedasticity (Ito & Schull, 1964). When larger samples are selected from populations with smaller dispersion (the negative condition), T^2 is liberal, and power is higher than it would be under conditions of homoscedasticity.

The effect of heteroscedasticity depends on the relationship between the sample sizes (n_1, n_2) and the eigenvalues (θ_j, *j*=1,...,*p*) (Algina, Oshima & Tang, 1991). When the eigenvalues are all equal, then the covariance matrices are related by the equation $\Sigma_1 = d^2\Sigma_2$, where $d^2 = \theta^{-1}$. Holloway and Dunn (1967) concluded that when eigenvalues are equal, covariance matrices are unequal

and sample sizes are unequal, T^2 may either be very liberal, or may be conservative with very low power. Changing these conditions to include equal sample sizes helped to control the Type I error rate in the liberal case, but did not increase the power to a useful level in the conservative case.

When all assumptions are met, T^2 is the uniformly most powerful test which is invariant to affine transformations (Anderson, 1958; Hakstian, Roed, & Lind, 1979; Hsu, 1938b; Olson, 1974). For a fixed total sample size, power is maximized when sample sizes are equal. Holloway and Dunn (1967) found that for fixed sample size, the power declined as the number of dependent variables increased.

When comparing Hotelling's T^2 to its alternatives (the James first- and second-order tests, Johansen's test, Kim's test, and Yao's test) under conditions of multivariate normality and heteroscedasticity, the following can be said:

(a) Yao's test outperforms Hotelling's T^2, particularly when sample sizes are unequal (Algina & Tang, 1988).

(b) The James first- and second-order tests, Johansen's test, and Yao's test have similar actual Type I error rates (Algina, Oshima & Tang, 1991; Subrahmaniam & Subrahmaniam, 1973). This, together with (1), implies that the James's tests and Johansen's test outperform Hotelling's T^2.

(c) The James second-order test, Johansen's test, and Yao's test control actual Type I error better than the James first-order test (Algina, Oshima & Tang, 1991).

(d) The ratio of the smaller sample size to p, the number of dependent variables, affects actual Type I error rates for the James tests, Johansen's test, and Yao's

test (Algina & Tang, 1988; Lin, 1991), but not for Kim's test (Kim, 1992). When $\min(n_1, n_2) / p < 4$, actual Type I error rates for the James, Johansen, and Yao tests are too high.

(e) When $\min(n_1, n_2) / p$ is small, Kim's test controls the actual Type I error rate better than Yao's test (Kim, 1992), and thus better than the James and Johansen tests.

(f) When $\min(n_1, n_2) / p > 4$, Johansen's test is adequate (Coombs & Algina, 1996-b).

Christensen and Rencher (1995) performed a limited simulation involving the above alternatives plus the Bennett test, the Hwang-Paulson test, and the Nel-van der Merwe test, using data sampled from multivariate normal populations, with unequal covariance matrices. They found that Bennett's test, Kim's test and Nel and van der Merwe's test were best at controlling Type I error.

Power investigations for alternatives to Hotelling's T^2 have been more limited. Subrahmaniam and Subrahmaniam (1973) looked at the Yao, Bennett, and James first-order tests under conditions of multivariate normality and unequal covariance matrices. They found that Bennett's test had low power. The James first-order test had the highest power, but the power of Yao's test was similar. The power of all three tests declined as the number of dependent variables (*p*) increased. Kim (1992) compared the Kim and Yao tests, again under conditions of multivariate normality and unequal covariance matrices, and found the tests to have similar power and control of Type I error. Yao's test had a slight power advantage in the positive condition, while Kim's test had a slight

power advantage in the negative condition. Christensen and Rencher (1995) compared power for Hotelling's T^2 with the Bennett, James first-order, Yao, Johansen, Nel-van der Merwe, Hwang-Paulson, and Kim tests. They found that Kim's test had both high power and a conservative Type I error rate. The Nel-van der Merwe test performed almost as well as Kim's test and is computationally simpler.

Performance of several alternatives to Hotelling's T^2 under conditions of non-normality was examined by Algina, Oshima and Tang (1991). They found that the James first-order, James second-order, Johansen and Yao tests have elevated Type I error rates for skewed distributions. The tests tended to be conservative for long-tailed symmetric distributions, and liberal for short-tailed symmetric distributions. The magnitude of the difference was dependent on the difference between sample sizes as well as the degree of heteroscedasticity and the degree of skewness. Everitt (1979) found that T^2 became more conservative as populations became more skewed, while Mardia (1975) found that that the effect of skewness was minimized with equal sample sizes.

Comparisons of Performance for MANOVA and Alternatives

Unlike the univariate two-group, univariate G-group, and multivariate two-group cases, none of the four MANOVA criteria (Roy's largest root, Hotelling-Lawley trace, Pillai-Bartlett trace, and Wilks's likelihood ratio) is uniformly the most powerful. When evaluating MANOVA and its alternatives, we therefore have the additional problem of deciding which of the four MANOVA criteria is best for a given situation. The literature suggests that when all MANOVA

assumptions are met, the following can be said about power for Roy's largest root (*R*), the Hotelling-Lawley trace (*U*), the Pillai-Bartlett trace (*V*), and Wilks's likelihood ratio (*L*):

(a) If population differences are concentrated along a single dimension, or dependent variable (this is known as concentrated noncentrality), then power(*R*) > power(*U*) > power(*L*) > power(*V*) (Olson, 1974; Pillai & Jayachandran, 1967; Schatzoff, 1966).

(b) If population differences are diffused over several dimensions, or dependent variables (this is known as diffuse noncentrality), then the order is reversed: power(*V*) > power(*L*) > power(*U*) > power(*R*) (Olson, 1974; Schatzoff, 1966).

(c) *U*, *L* and *V* are asymptotically equivalent (Olson, 1974).

(d) *U*, *L* and *V* are superior to *R*, because the power of *R* declines appreciably under diffuse noncentrality (Olson, 1974; Schatzoff, 1966). The possibility of diffuse noncentrality (i.e., differences on more than one dependent variable) is what the researcher hopes to uncover by choosing a multivariate test, so a loss of power under this condition is serious.

(e) All MANOVA criteria have power problems with small group sizes, even for moderate effect sizes (Stevens, 1980). Since small to medium effect sizes are common in social science research (Becker, 1987; Cohen, 1988; Stevens, 1996), this is also a serious concern.

Under conditions of heteroscedasticity, the behavior of MANOVA is similar to that of the ANOVA *F* test, as has been documented by a number of studies (Ito & Schull, 1964; Korin, 1972; Olson, 1974; Pillai & Sudjana, 1975; Tang &

Algina, 1993). Even with equal sample sizes, actual Type I error rates can be inflated with unequal covariance matrices, and the MANOVA tests can be either liberal or conservative when sample sizes are unequal, depending on whether the larger samples come from populations with large or small variances. Olson (1974) and Elliot and Barcikowski (1994) both came to the conclusion that the Pillai-Bartlett (*V*) test was the best of the four MANOVA criteria at controlling Type I error with unequal covariance matrices.

The effect of assumption violations on power has not been as extensively studied. Olson (1974) studied the effects of both nonnormality and heteroscedasticity for MANOVA criteria under a variety of conditions. He used the term "contamination" to refer to localized assumption violations, such as a particular dependent variable and group with nonnormality. This is a particularly useful terminology for MANOVA since we have both multiple groups and multiple dependent variables. Olson found that contamination decreased power (and affected Type I error rates, as noted above), but if noncentrality occurred in a noncontaminated group or variable, power was maintained. He also found that kurtosis decreased power for all four MANOVA tests, and that unequal covariance matrices caused all power curves to be rather flat.

Olson (1976; 1979) and Stevens (1979, 1996) have disagreed over which of the four MANOVA criteria is preferred for general use. Olson recommended *V* (Pillai-Bartlett) because (a) *V* was most robust under conditions of heteroscedasticity; (b) *V* was the least conservative when sampling from platykurtic populations (Ito, 1969; Ito & Schull, 1964; Korin, 1972; Mardia, 1971;

Olson, 1974); and (c) the power of *V* was adequate, even though it was the least powerful of the four MANOVA criteria under heteroscedasticity. Stevens criticized Olson's conclusion, in part, because he considered only equal-sized samples, and because the heteroscedasticity conditions used were extreme and not likely to occur in practice. Stevens conceded the superiority of *V* under conditions of diffuse noncentrality, but pointed out that for concentrated noncentrality with unequal covariance matrices, the actual Type I error rates for *U*, *V* and *L* are similar. Since *U* and *L* have slightly greater power, he recommended their use over *V*. Olson's rejoinder was that since *V* is clearly superior in one area (diffuse noncentrality), and *U*, *V* and *L* are all similar (with slight differences) in other areas, *V* is still the best choice.

The above discussion points out that there is no easy answer to the question, "Which MANOVA test is superior?" Add to this the observation by Coombs, Algina, and Oltman (1996) that for an experiment with three groups and four dependent variables, 20 independent assumptions are actually being made about the equality of pairs of covariance matrices. This suggests that in practice, the assumption of homogeneity of covariance matrices is untenable. Alternatives which do not require this assumption should then assume greater importance. It should also be noted that common statistical packages will automatically give the researcher all four MANOVA criteria. If some of these tests show significance while others do not, the researcher is given a difficult decision. Are some tests not significant because of low power due to assumption violations, or are others significant because of conservative Type I error rates due to assumption

violations? Perhaps instead of picking a "best" alternative among the four, we should advise researchers that if some but not all of the MANOVA tests show significance, then proceed very cautiously.

When sampling from multivariate normal populations with equal sample sizes and unequal covariance matrices, the following can be said in comparing the Pillai-Bartlett test with the Coombs-Algina tests, James first- and second-order tests, and the Johansen test:

(a) The James first-order and Johansen tests tend to be somewhat liberal, with the Johansen test giving better control of Type I error (Tang & Algina, 1993).

(b) The James second-order test tends to be somewhat conservative (Tang & Algina, 1993).

(c) The Pillai-Bartlett test can be liberal, with the degree of difference between the actual and nominal Type I error rate depending on the degree of difference between the population covariance matrices (Olson, 1974; Tang & Algina, 1993).

(d) When sample sizes are sufficiently large, the Johansen test performs better than the Pillai-Bartlett test (Tang & Algina, 1993).

(e) When sample sizes are too small, Johansen's test has a greatly inflated actual Type I error rate, but the James second-order test and the Coombs-Algina U^* appear to control Type I error fairly well under this condition (Coombs & Algina, 1996-b; Tang & Algina, 1993).

(f) With sufficiently large sample sizes, Johansen's test controls Type I error rate better than the Coombs-Algina tests or the James second-order test (Coombs &

Algina, 1996-b; Tang & Algina, 1993), but since the James second-order test tends to be slightly conservative, some researchers may favor it.

When sampling from populations with multivariate normality, with unequal covariance matrices and unequal sample sizes, the literature suggests the following:

(a) The Pillai-Bartlett test is not recommended, because its actual Type I error rate can be substantially greater than or less than α (Tang & Algina, 1993).

(b) The Johansen test performs better than the James first-order test (Tang & Algina, 1993).

(c) When using the Johansen test, the researcher should have a ratio of at least 3 1/3 for the minimum sample size to the number of independent variables with G = 3 groups, and a ratio of at least 4 2/3 with G = 6 groups, in order to achieve adequate control of Type I error rate. If the ratio of minimum sample size to number of independent variables is smaller than 4, then the Coombs-Algina U^* and James' second-order test appear to be the best choices (Coombs & Algina, 1996-b; Tang & Algina, 1993).

The MANOVA tests and their alternatives were all developed under the assumption of multivariate normality. If this assumption is violated, the Coombs-Algina tests appear to be the best at controlling Type I error rates. Oltman (1996) studied the performance of the Pillai-Bartlett test, Johansen's test, and the Coombs-Algina tests when sampling from populations with extreme skew (an exponential distribution). She found that the Coombs-Algina tests were able to control Type I error rates well under these conditions, but their power was too low

to be useful. Neither Johansen's test nor the Pillai-Bartlett test controlled Type I error rates well when sampling from exponential distributions. This suggests that for distributions which are skewed (but not as strongly skewed as the exponential distribution), the Coombs-Algina tests will probably control Type I error well, and at some threshold, may also offer adequate power.

The Independence Assumption

The assumption that each observation is independent of all other observations is rarely studied in Monte Carlo simulations. However, as Glass, Peckham, and Sanders (1972) observe in their review of ANOVA studies, "The violation of the independence assumption which we shall not discuss ... is far more serious than the violation of the assumptions which we will discuss" (p. 242). In general, the advice given to researchers is that a good research design will avoid problems with the independence assumption. In many cases, there are formalized patterns of dependence that may be incorporated into the statistical analysis, such as repeated measures. In other cases, such as studies of teaching methods using intact classrooms, the unit of analysis may be changed from the student to the classroom (Pedhazur, 1982). However, these techniques do not cover all potential problems.

The small body of literature dealing with dependence focuses on ANOVA designs. The formulas in this section, unless otherwise noted, are all applicable to ANOVA. Dependence due to groups is usually measured by the intraclass correlation. There are a number of different estimates of intraclass correlations (Lahey, Downey &Saal; 1983), but the most common is defined as follows:

$$\frac{MS_b - MS_w}{MS_b + MS_w(n-1)},$$

where MS_b and MS_w are as defined on page 13, and *n* is the sample size for each group.

Shavelson (1988) notes that the intraclass correlation provides a measure of the extent to which within-group variability is small relative to between-group variability. It is at its maximum when scores within groups are identical and the group means differ among one another.

Kenny and Judd (1986) discuss three characteristic patterns of dependence likely to occur in social science research: dependence due to groups; dependence due to sequence; and dependence due to space. Dependence due to sequence is caused when observations taken over time are not independent due to cyclical patterns. As an example, the number of calories consumed per day may be greater on the weekends when people are at home all day. This form of dependence may in many cases fit a repeated measures or time series design. Dependence due to space is most often found due to the "closest neighbor" effect, where observations taken from the same block or neighborhood are related. This relationship diminishes with distance.

This study will simulate conditions of dependence due to groups, which occurs when subgroups within treatment groups are related. They may be informal groups of friends, or the groups may actually be part of the design, as is the case when comparing the effects of group discussion and lecture formats in the classroom. In the latter case, the design itself needs to be changed. The remedies suggested include the quasi-*F* test (Myers, DiCecco & Lorch, 1981).

Pavur and Nath (1984) detail the application of a constant multiplier to correct for correlation within groups. This solution will work perfectly only when the within-group correlation is uniform.

One of the more curious properties of dependence of observations is that its effect on Type I error rate increases as sample sizes increase. (Recall that for violations of the assumption of normality, large sample sizes mitigate the distortions of Type I error.) Scariano and Davenport (1987) demonstrate that as the sample size of each group approaches infinity, the true Type I error rate for dependent samples approaches 1.0. Even for moderate correlations such as $\rho = .3$ and for 2 groups, the Type I error rate for n=30 is .5928; for n=100 it is .7662 (see Table 2).

What kind of intraclass correlation might be expected in an educational setting, assuming that we have no formal groups? Though group relationships should be expected, their effects may be small and may work in different directions. Further, groups may exist across treatment boundaries as well as within them. Kenny and Judd (1986) demonstrate the influence of intraclass correlation and intergroup correlation on the sampling distributions of MS_b and MS_w. The means of the sampling distributions of MS_b and MS_w are, respectively:

$$E(MS_b) = \sigma^2\left[1 + (n-1)\rho_w - n\rho_b\right] + n\sigma_\alpha^2$$
$$E(MS_w) = \sigma^2\left(1 - \rho_w\right)$$

where ρ_w is the average dependence among pairs of observations in the same treatment group, ρ_b is the average dependence among pairs of observations

Table 2. Actual Type I Error Rates for Different Intraclass Correlations (nominal Type I error rate = .05)

Num. Grps	Grp. Size	ICC= .00	ICC= .01	ICC= .10	ICC= .30	ICC= .50	ICC= .99	ICC= .70	ICC= .90	ICC= .95
2	3	.0500	.0522	.0740	.1402	.2374	.8800	.3819	.6275	.7339
	10	.0500	.0606	.1654	.3729	.5344	.9475	.6752	.8282	.8809
	30	.0500	.0848	.3402	.5928	.7205	.9708	.8131	.9036	.9335
	100	.0500	.1658	.5716	.7662	.8446	.9842	.8976	.9477	.9640
3	3	.0500	.0529	.0837	.1866	.3430	.9829	.5585	.8367	.9163
	10	.0500	.0641	.2227	.5379	.7397	.9966	.8718	.9639	.9826
	30	.0500	.0985	.4917	.7999	.9049	.9990	.9573	.9886	.9946
	100	.0500	.2236	.7791	.9333	.9705	.9997	.9872	.9966	.9984
5	3	.0500	.0540	.0997	.2684	.5149	.9997	.7808	.9704	.9923
	10	.0500	.0692	.3151	.7446	.9175	1.0000	.9798	.9984	.9996
	30	.0500	.1192	.6908	.9506	.9888	1.0000	.9977	.9998	1.0000
	100	.0500	.3147	.9397	.9945	.9989	1.0000	.9998	1.0000	1.0000
10	3	.0500	.0560	.1323	.4396	.7837	1.0000	.9664	.9997	1.0000
	10	.0500	.0783	.4945	.9439	.9957	1.0000	.9998	1.0000	1.0000

30	.0500	.1594	.9119	.9986	1.000	1.000	1.000	1.000	1.000
					0	0	0	0	0
100	.0500	.4892	.9978	1.000	1.000	1.000	1.000	1.000	1.000
				0	0	0	0	0	0

which are not in the same treatment group, σ^2 is the population variance for each treatment group, and σ_α^2 is the variance of the treatment effect.

Besides its effects on the expected value of mean squares, dependence can effect statistical results in two other ways. Box (1954a, 1954b) showed that dependence can increase the variability of mean square estimates. Kenny and Judd (1986) also point out that when observations are dependent, MS_b and MS_w may be correlated. Because of this correlation, the computed *F* ratio may not be distributed as *F*.

Chapter 3

Method

This study will examine the robustness and power of MANOVA and several MANOVA alternatives under a set of simulated conditions. It will build upon and add to conditions already studied in previous Monte Carlo simulations. Glass, Peckham, and Sanders (1972) provide a template for Monte Carlo studies of the robustness and power of statistical tests, which will be followed here:

1) Given a value for α and values for necessary degrees of freedom, the critical values of the statistical tests in the simulation are found. This value of α is called the nominal α.
2) Simulated data exhibiting the desired characteristics are randomly generated by computer program. The desired characteristics include the specified deviations from assumptions plus a given effect size (a fixed difference between group means). An effect size of zero (no difference between groups) will be included in the simulated characteristics, in order to get an estimate of Type I error under the specified deviations.
3) Through an iterative loop, simulated data are repeatedly generated and evaluated using the chosen statistical tests. At the end of the simulation, an estimate of α is calculated for each test as the percentage of times the critical value of the test was exceeded by the simulated data.
4) For an effect size of zero, the estimated α is compared to the nominal α. An actual value greater than the nominal value implies a liberal test (the test

found a difference more often than would be expected); an actual value less than the nominal value implies a conservative test.

5) For effect sizes greater than zero, the estimated α values are compared among the different statistical tests used, to get an estimate of the relative power of those tests.

Our ultimate goal is to identify the statistical test which, while holding the Type I error rate close to the nominal α level, maximizes power.

Olson (1974) notes that "One of the first problems in a study of MANOVA robustness arises from the apparently unlimited number of ways in which the assumptions of normality and covariance homogeneity can be violated" (p. 895). This is also true of violations of the assumption of independence. Another observation is that in Monte Carlo studies, it often makes more sense to vary operands which are directly related to the generation of data characteristics, rather than those which are related to output characteristics. For example, Olson's (1974) violations of the assumption of covariance homogeneity are based on a constant multiplier *D*, which is used to selectively multiply data from some groups but not others. His design factors are based on *D*, rather than on the variance of the different groups, because it is clear that increasing *D* will increase the heterogeneity of variance.

Design Factors

Statistical Tests (T):

The statistical tests which will be compared in this study include the Pillai-Bartlett *V*, the Johansen *J*, the Coombs-Algina U_1^*, the Coombs-Algina U_2^*, the Coombs-Algina L^*, and the Coombs-Algina V^* tests.

Dependence of Observations:

This study will examine violations of the assumption of independence of observations. Since even moderate violations of this assumption are known to cause severe distortions in Type I error rate and power for univariate procedures (Glass, Peckham & Sanders, 1972; Scariano & Davenport, 1987), only small and moderate violations will be examined here. The Type I error rate and power of the multivariate tests being considered in this study have not been previously investigated for performance under violations of the independence assumption. Because of this, it will be necessary to operationally define how to generate data with different levels of dependence.

This study will simulate dependence due to groups. To be more precise, we will look at what happens when there is correlation in subgroups within treatment groups. Subgroups of varying sizes will be contained within treatment groups, and the dependence will come from correlated errors within these subgroups. Subgroup sizes of 2, 3, 4 and 6 will be simulated. These subgroup sizes were picked to represent the smallest, and thus most likely, informal groups that might be found in a classroom where close friends may study together.

In order to control the size of the treatment groups, we will then vary the number of subgroups per treatment group. The subgroup sizes of 2, 3, 4 and 6 will correspond to, respectively, 6, 4, 3 and 2 numbers of subgroups per treatment group, since all three combinations give us treatment groups of size 12. We can then vary the numbers of subgroups from 12 to 8 to 6 to 4, corresponding to subgroup sizes of 2, 3, 4, and 6, respectively, to give treatment groups of size 24.

We will also vary the degree of correlation among the errors found within each subgroup. This corresponds to the degree to which we expect that working within the subgroup would affect performance on our dependent variables. A high degree of correlation between the errors within each subgroup results in lower variability within each subgroup, reducing the total variability within each treatment group. Recall that each person's score on the set of dependent variables can be thought of as the sum of the group mean plus error (individual variability). In a typical Monte Carlo study, without simulating dependence of observations, we would have

$$\mathbf{y}_{ij} = \mu_i + \varepsilon_{ij}$$

where *i* is the treatment group, and *j* is the observation within the treatment group. The error term is randomly generated for each observation. We will take this one step further and break down the error component into both fixed and random components within each subgroup. If subscript *k* represents the subgroup, we have

$$\mathbf{y}_{ijk} = \mu_i + \delta_k + \varepsilon_{ijk}$$

where μ varies with each treatment group, δ varies with each subgroup, and ε varies with each observation. Note that the treatment means will be determined by our effect size, whereas the other two components will be randomly generated. Now we wish to find a value Rho for which the following are true:

a) increases in Rho will increase the dependence among observations within treatment groups;
b) the correlation between any two error components within the same subgroup is equal to Rho;
c) the correlation between any two error components in different subgroups is zero;
d) the expected value of each error component is zero; and
e) the standard deviation of each error component is one.

Appendix A gives the details for generating observations based on values of Rho which satisfy the above conditions.

The levels of dependence considered in this study will correspond to values of Rho = 0.0, 0.01, 0.2, 0.4, and 0.6.

Number of Groups (G):

A review of related studies suggests that researchers have varied the number of groups of the independent variables in Monte Carlo studies of MANOVA from two to ten, with three and six groups being the most common conditions. These studies include Korin (1992), Tang (1989), Coombs (1993), Brown & Forsythe (1974), Kohr & Games (1974), Clinch & Keselman (1982), Tomarken & Serlin (1986), Wilcox, Charlin & Thompson (1986), Wilcox (1988,

1989), Dijkstra & Werter (1981), and Olson (1974). This study will use two and three groups.

Number of Dependent Variables (P):

A review of related studies (Korin 1992; Tang 1989; Coombs 1993; Brown & Forsythe 1974; Kohr & Games 1974; Clinch & Keselman 1982; Tomarken & Serlin 1986; Wilcox, Charlin & Thompson 1986; Wilcox 1988, 1989; Dijkstra & Werter 1981; and Olson ;1974) suggests that the most common choices for the number of dependent variables are three and six. However, Oltman (1996) found that this factor was not important in explaining differences in rejection rates (under different conditions than those which will be studied here). This study will look at both two and three dependent variables.

Effect Size

The effect size is a measure of the degree of difference between groups. A simple formula for effect size in the univariate case with two groups is as follows:

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma},$$

where μ_1, μ_2 are the population means of two different groups, and σ is the population variance.

In the multivariate two-group case, the extension of this formula is called the Mahalanobis distance, given by

$$D^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$$

In practice, the population means and covariance matrices are estimated by their sample counterparts.

Stevens (1992) notes that for the *G*-group multivariate case, where *G*>2, most Monte Carlo studies have assessed effect size through measures which use eigenvalues, and hence make it difficult for the reader to determine the actual differences between means which were used in the study.

In power studies, different effect sizes are examined to assess the relative power of different statistical tests. With an effect size of zero, we are testing only the Type I error rate. This study will examine effect sizes corresponding to no effect, a moderate effect and a large effect. The exact values will be determined through the noncentrality function, discussed below.

Type of Noncentrality

In order to determine the relative power of the statistical tests under consideration, we will look at both the noncentrality structure and the noncentrality function (Olson, 1974).

In a power analysis, we set the degree of difference, or effect size, between population groups on the different dependent variables, then look to see how well our tests can detect that difference. In *G*-group MANOVA, there are *S=min(G,P)* eigenvalues, with corresponding eigenvectors which define a linear combination of the dependent variables. These linear combinations are the discriminant functions.

The noncentrality structure refers to the degree that differences present are "concentrated" on one of the discriminant functions, or "diffused" among several discriminant functions. Concentrated noncentrality may be simulated by fixing a difference on one group and on one dependent variable only (we will

arbitrarily pick the first one). Diffuse noncentrality may be simulated by fixing the same difference among all groups on all dependent variables.

The exact amount of the difference may be set by the noncentrality function. Schatzoff (1966) defined the noncentrality function as the trace of matrix **G**, tr(**G**),

where

$$\mathbf{G} = \mathbf{H}\mathbf{V}^{-1},$$

V is the population covariance matrix, and

$$\mathbf{H} = \sum_{i=1}^{G} n_i(\mu_i - \mu)(\mu_i - \mu)',$$

where μ_i is the population mean vector for the *i*th group, μ is the grand mean vector, and n_i is the sample size in the *i*th group.

In practice, the noncentrality parameter tr(**G**) can be estimated by (*N* - *G*) times the Hotelling-Lawley trace. Olson (1974) used four levels in his study: 0, 10; 40; and 90. These correspond roughly to no effect, small effect, medium effect, and large effect.

Having set the desired values of the noncentrality parameter, it remains only to fix the mean values which will give us the desired noncentrality parameter and structures. For concentrated noncentrality, we shall let the first group differ from the other (*G-1*) groups by setting the (1,1) element of the mean vector of the first population to *Gc*, where *c* is determined by solving the following equation (Olson, 1974):

$$tr(\mathbf{G}) = nG(G-1)c^2.$$

The other values of the mean vector are zero, as are all of the values of the mean vectors of the other (*G-1*) groups.

For diffuse noncentrality, we will set the *i*th group of the *i*th dependent variable to *Gc*, where *c* is determined by

$$tr(\mathbf{G}) = (p-1)nG^2c^2 + nG(G-p)c^2$$

when *G>p*, and

$$tr(\mathbf{G}) = (G-1)nG^2c^2$$

when *G<=p*. The other values of the mean vectors will be zero.

The values of *tr(**G**)* will be set to 0, 40 and 90.

Design Layout

Each of the six statistical tests to be compared here will be performed on 5 x 4 x 2 x 2 x 2 x 3 x 2 = 960 different condition combinations representing the levels of conditions specified above (Rho x Nspsub x Nsubpgr x G x P x Effect x NCP). Each of these conditions will be repeated five times.

Simulation Procedure

The simulation will be conducted as 960 x 5 = 4,800 separate analyses, with 1,000 replications per condition. For each condition, the performance of the Pillai-Bartlett *V*, the Johansen *J*, the Coombs-Algina U_1^*, the Coombs-Algina U_2^*, the Coombs-Algina L^*, and the Coombs-Algina V^* tests will be evaluated using the generated data, following the template presented at the beginning of this chapter. The proportion of the 1,000 replications which yield significant results at α =.05 will be tabulated; these will serve as estimates of the rejection rates of the tests for the various condition combinations.

The computer program to perform the simulation will be written in the SAS language, using PROC IML (Interactive Matrix Language) to accommodate the tests (the Johansen and Coombs-Algina tests) which are not available as options in regular SAS procedures. Existing programs from previous studies on the same group of statistical tests will be modified to create the correct combination of conditions.

Chapter 4

Results

In this chapter the rejection rates from the simulation are presented and discussed for the combinations of conditions under study.

Rejection rates were analyzed using a split-plot analysis of variance model. One within factor (test criterion) and seven between factors (type of noncentrality, effect size, number of groups, number of dependent variables, size of treatment group, size of subgroups, and degree of dependence) were included in the model, along with all possible interactions. Each combination of conditions was replicated five times so that within cell variation could be assessed for the model.

Not surprisingly, given the large sample sizes (1,000 repetitions per condition), there were a number of statistically significant effects, including interactions of up to five dimensions. Interactions of more than three dimensions cannot be graphed using a single figure, and are difficult to interpret. The approach taken here will be to provide groups of graphs in order to examine dimensions which will not fit on one graph. Out of 56 five-way interaction terms, seven were statistically significant at alpha=.05. These significant five-way interactions together involved all of the eight independent variables in the model. There were a number of other significant interaction terms of lower dimension, and all eight main effects were significant.

One of the significant five-way interactions was Test x Rho x Nspsub x Nsubpgr x G, where Nspsub is the number of subjects per subgroup and

Nsubpgr is the number of subjects per treatment group. Recall that we are trying to discover how the estimated alpha levels found in our simulation compare to our nominal value of 0.05. The "rejection rate" refers to the proportion of times in our simulation that statistical significance was found for a given set of conditions. Figures 1 through 10 graph the rejection rates for some of the combinations of these five variables. For these figures, the number of dependent variables was held constant at P=2, and Effect Size was held constant at zero. In the two-group case, all four of the Coombs-Algina tests are equivalent (as are all four of the MANOVA criteria), so only one line is shown for the Coombs-Algina tests in Figures 1 through 8 (Coombs, 1993). We can see that (a) increases in Rho correspond to increases in rejection rate; (b) rejection rates increase more sharply (with increases in Rho) when the number of subjects per subgroup increases; (c) though the differences between the tests were statistically significant, their rejection rates were always within one percentage point; and (d) increasing the number of groups does increase the rate of increase in Rho (other factors held constant).

Figure 1
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Subjects per Subgroup=2, Effect Size=0, Group Size=12, Number of Groups=2, and Number of Dependent Variables=2

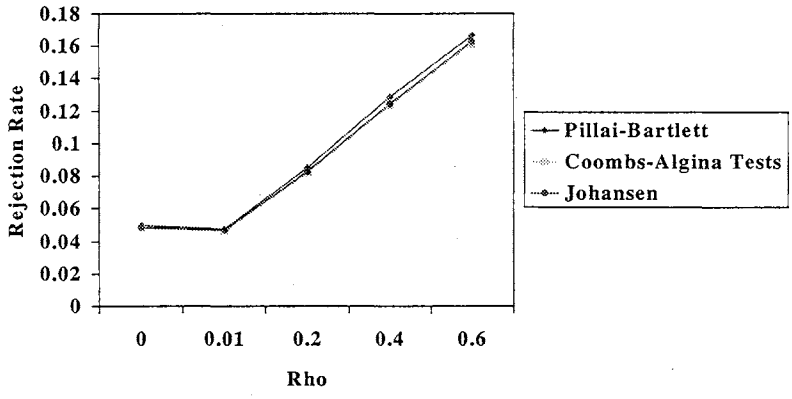


Figure 2
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Subjects per Subgroup=3, Effect Size=0, Group Size=12, Number of Groups=2, and Number of Dependent Variables=2

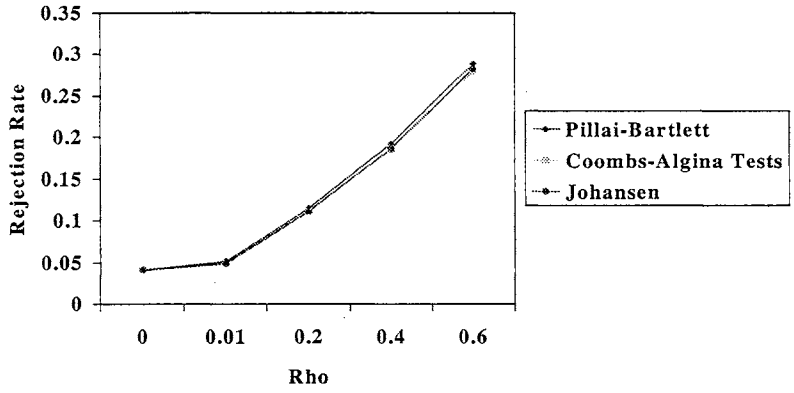


Figure 3
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests
With Subjects per Subgroup=4, Effect Size=0, Group Size=12, Number of Groups=2, and Number of Dependent Variables=2

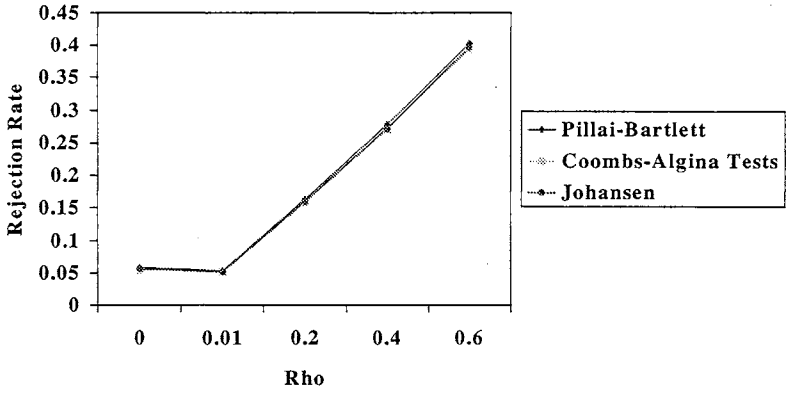


Figure 4
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests
With Subjects per Subgroup=6, Effect Size=0, Group Size=12, Number of Groups=2, and Number of Dependent Variables=2

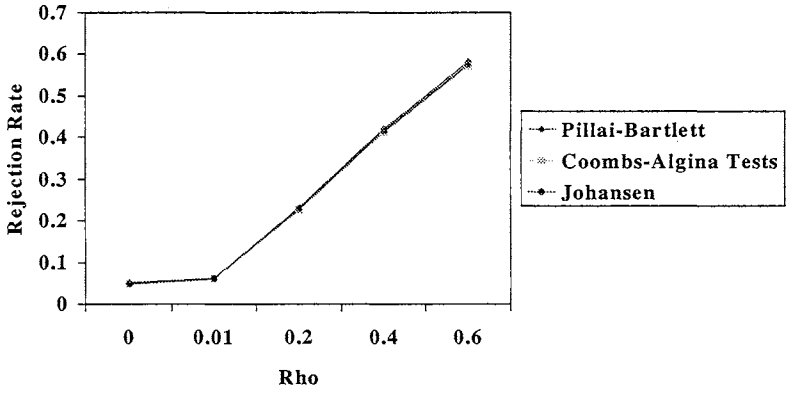


Figure 5
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Subjects per Subgroup=2, Effect Size=0, Group Size=24, Number of Groups=2, and Number of Dependent Variables=2

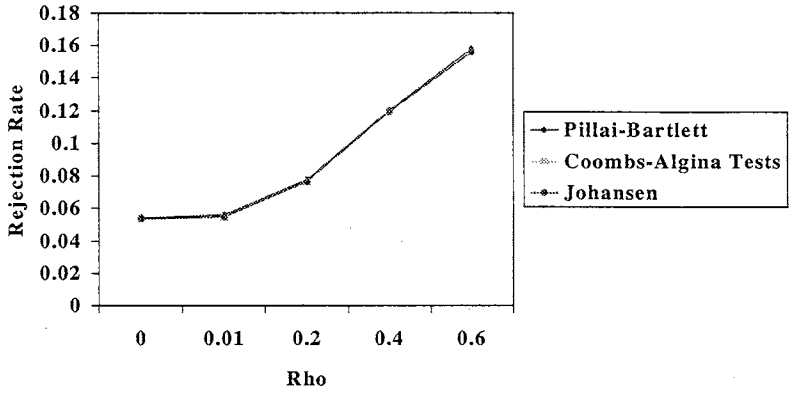


Figure 6
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Subjects per Subgroup=3, Effect Size=0, Group Size=24, Number of Groups=2, and Number of Dependent Variables=2

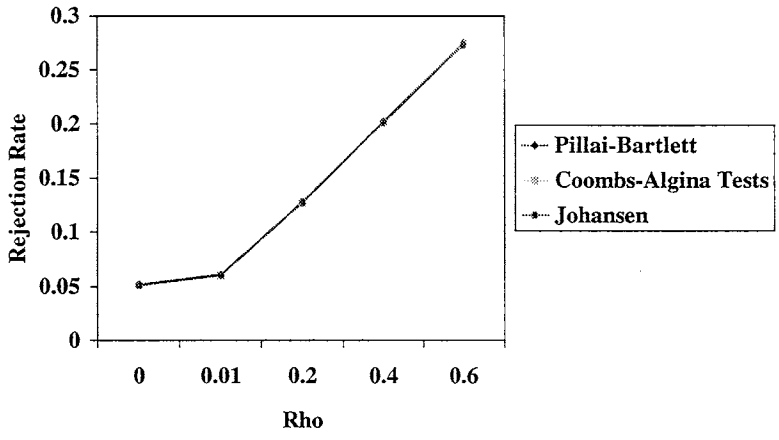


Figure 7
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Subjects per Subgroup=4, Effect Size=0, Group Size=24, Number of Groups=2, and Number of Dependent Variables=2

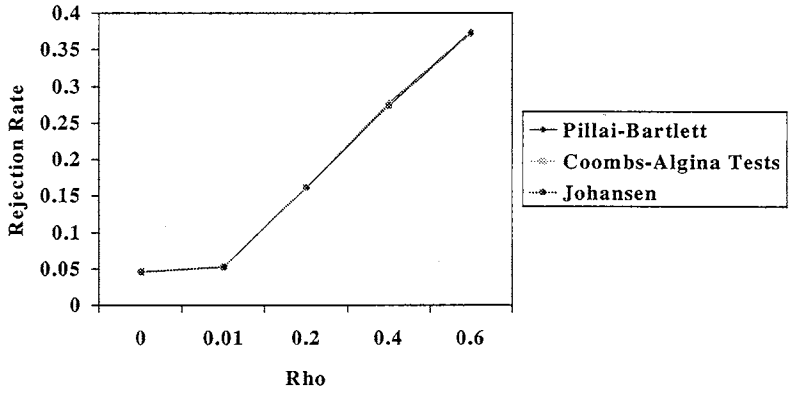


Figure 8
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Subjects per Subgroup=6, Effect Size=0, Group Size=24, Number of Groups=2, and Number of Dependent Variables=2

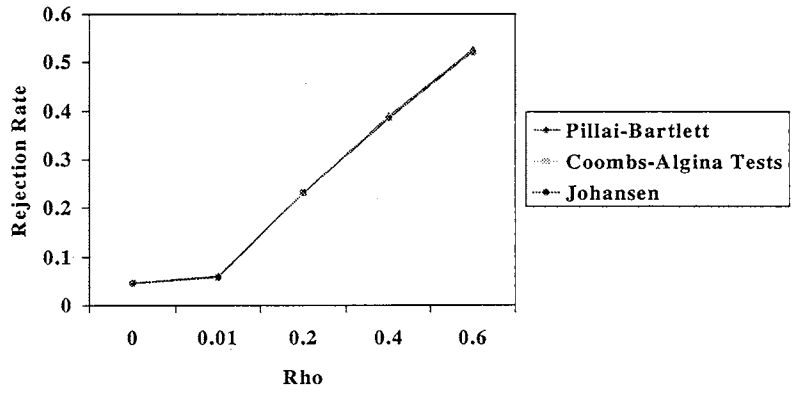


Figure 9
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Subjects per Subgroup=2, Effect Size=0, Group Size=12, Number of Groups=3, and Number of Dependent Variables=2

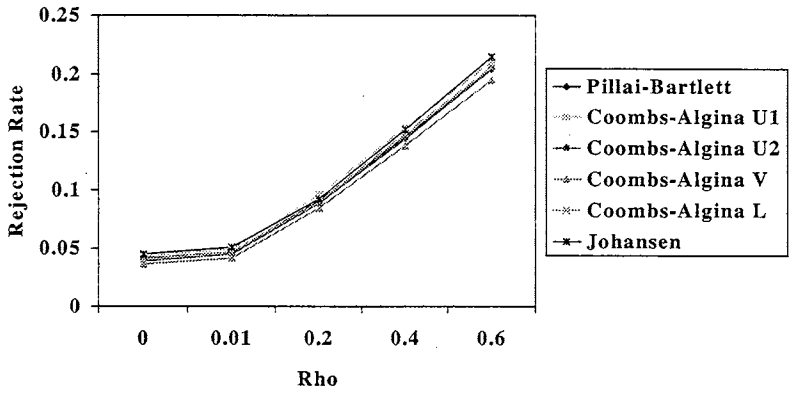
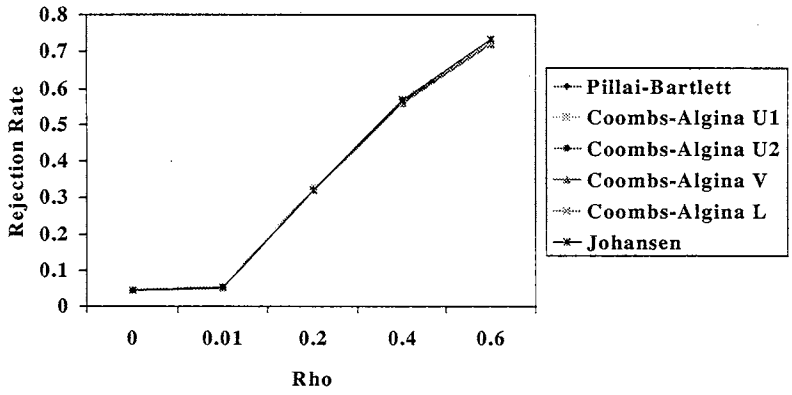


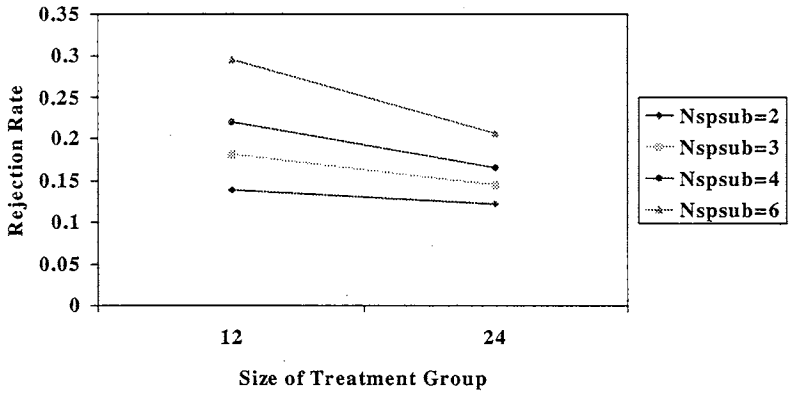
Figure 10
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Subjects per Subgroup=6, Effect Size=0, Group Size=24, Number of Groups=3, and Number of Dependent Variables=2



Surprisingly, when the treatment group size goes from 12 to 24 and Rho increases, the rate of increase in rejection rate goes down. For example, at Rho=0.6 and number of subjects per subgroup=6, the rejection rate is approximately 58% with a treatment group size of 12, while with a treatment group size of 24 the rejection rate is approximately 52%. This at first seems

counter-intuitive, since an increase in the treatment group size will increase the total sample size. Table 2 in Chapter 3 shows us that for a constant value of the intraclass correlation, increasing the sample size will increase the rejection rate. However, this study has been organized so that we are not varying the intraclass correlation directly. Figure 11 shows that, holding other factors constant, as the treatment group size goes from 12 to 24, the intraclass correlation actually drops slightly. This may be explained by noting that increasing the treatment group size, while leaving the number of subjects per subgroup the same, may dilute the effects of the reduction in variance.

Figure 11
Comparison of Intraclass Correlation
With Rho=0.6, Effect Size=0, Number of Groups=2, and Number of Dependent Variables=2



Another significant five-way interaction was Test x G x P x Nsubpgr x Rho. Figures 12 through 19 show that (a) the differences in rejection rates between tests are minimal (under 1%); (b) as the number of dependent variables increases, the rejection rate increases; (c) as the number of groups increases,

the rejection rate increases; (d) as the number of subjects per treatment group increases, the rejection rate decreases; and (e) as Rho increases, the rejection rate increases. In each of these figures, the number of subjects per subgroup was held constant at Nspsub=2, and the Effect Size was zero.

Figure 12
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

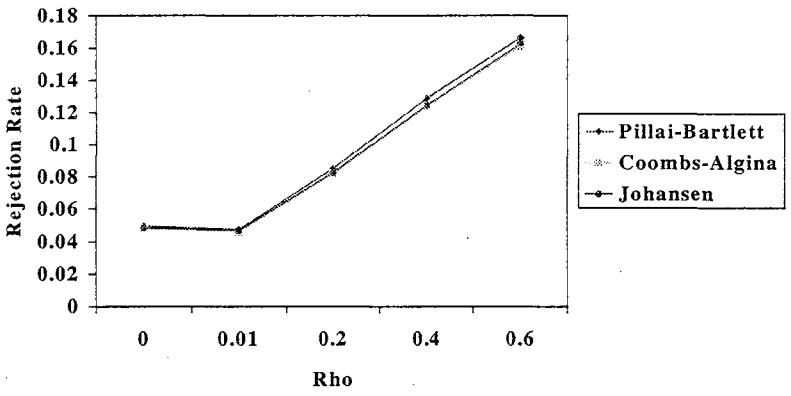


Figure 13
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24

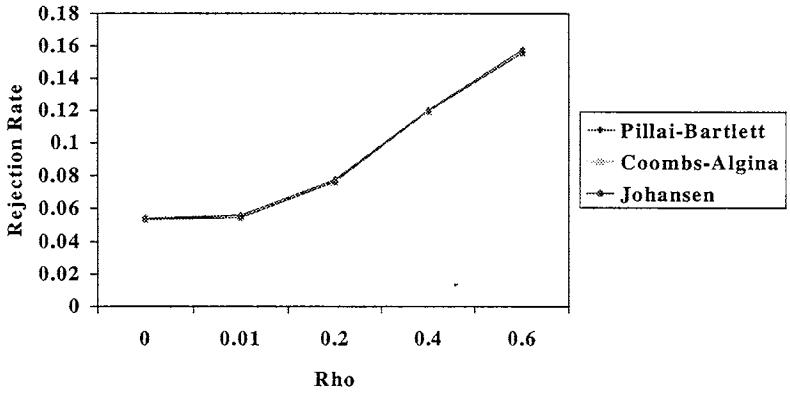


Figure 14
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=3 and Number of Subjects per Treatment Group=12

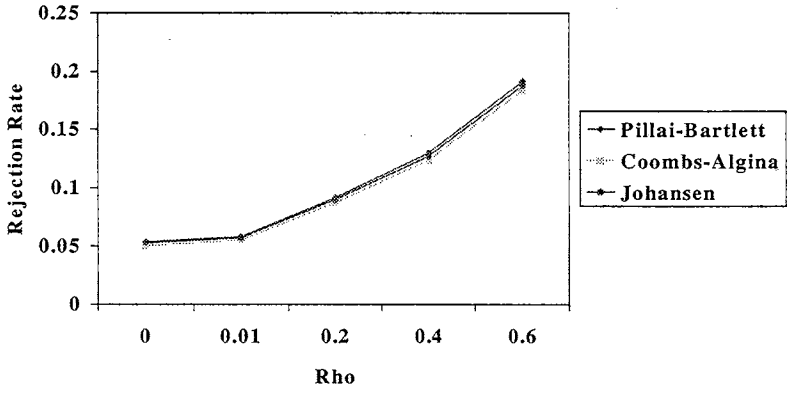


Figure 15
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=3 and Number of Subjects per Treatment Group=24

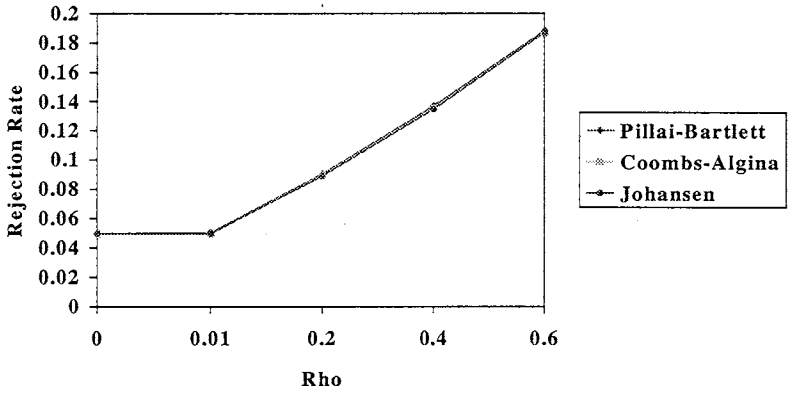


Figure 16
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

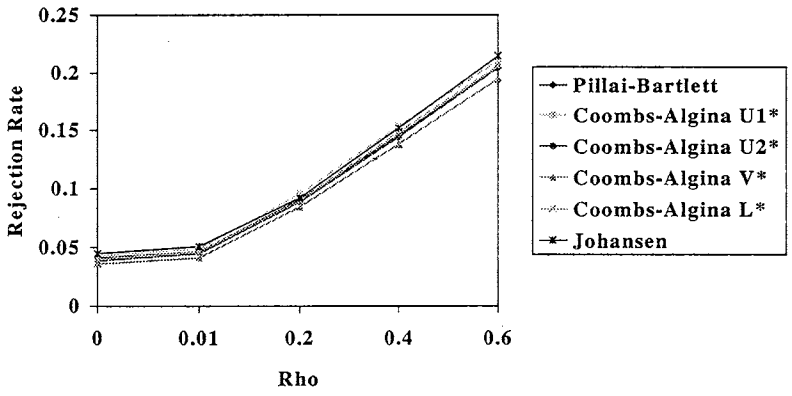


Figure 17
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24

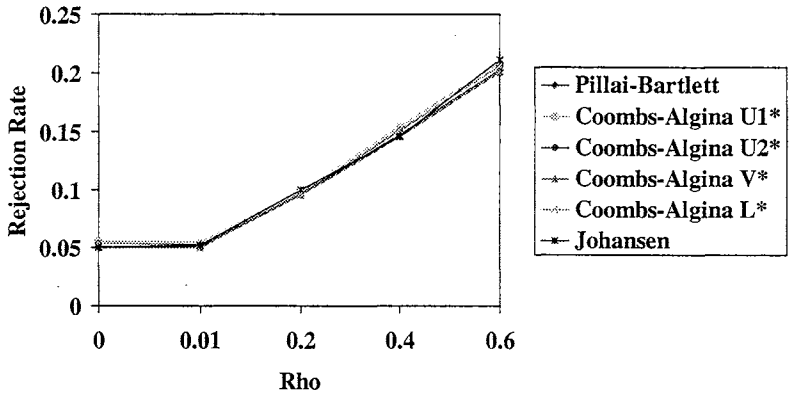


Figure 18
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=3, Number of Dependent Variables=3 and Number of Subjects per Treatment Group=12

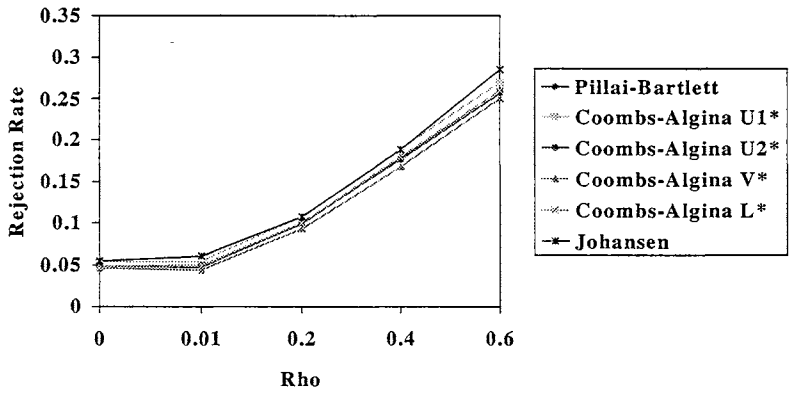
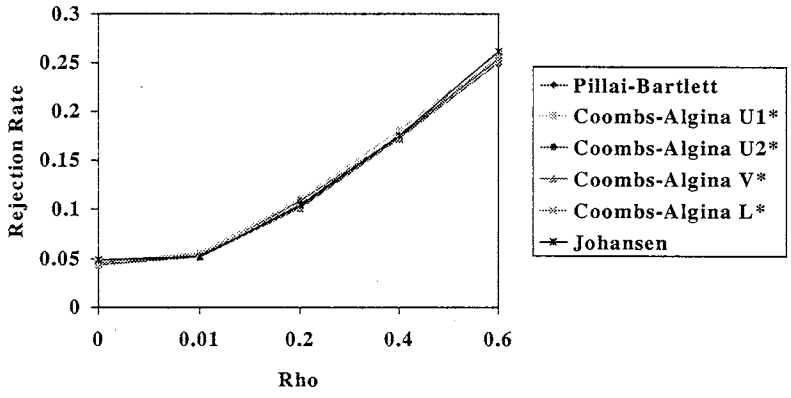


Figure 19
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=3, Number of Dependent Variables=3 and Number of Subjects per Treatment Group=24



A third significant five-way interaction was Test x G x P x Nspsub x Rho. Figures 20 through 27 show that (a) the differences in rejection rates between tests are minimal (under 1%); (b) as the number of dependent variables increases, the rejection rate increases; (c) as the number of groups increases, the rejection rate increases; (d) as the number of subjects per subgroup increases, the rejection rate increases; and (e) as Rho increases, the rejection rate increases. In each of these figures the number of subjects per treatment group was held constant at Nsubpgr=12, and the Effect Size was zero.

Figure 20
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Subgroup=2

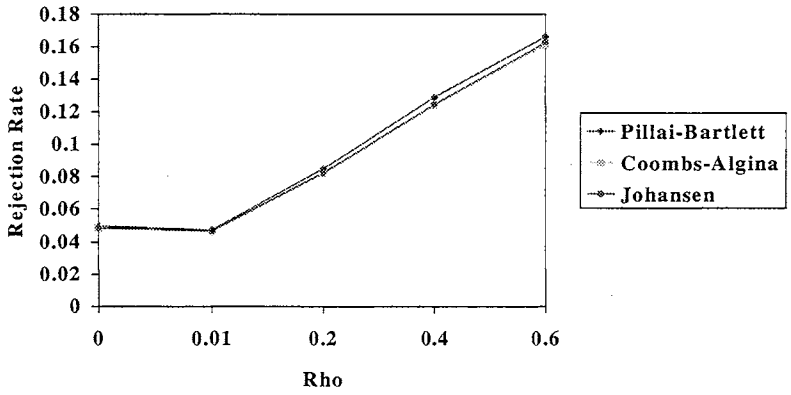


Figure 21
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Subgroup=6

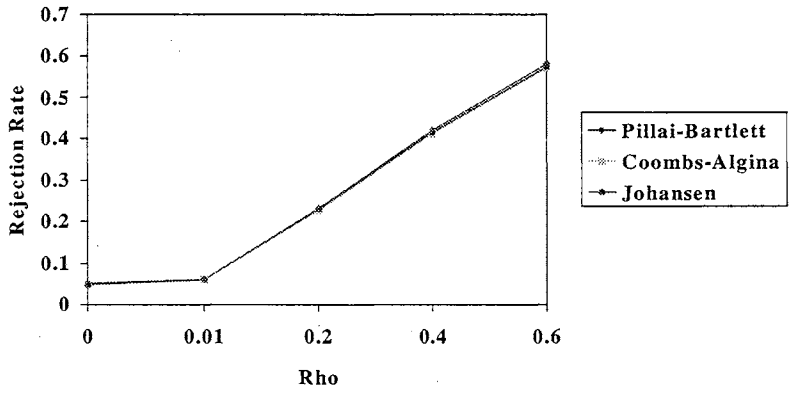


Figure 22
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=3 and Number of Subjects per Subgroup=2

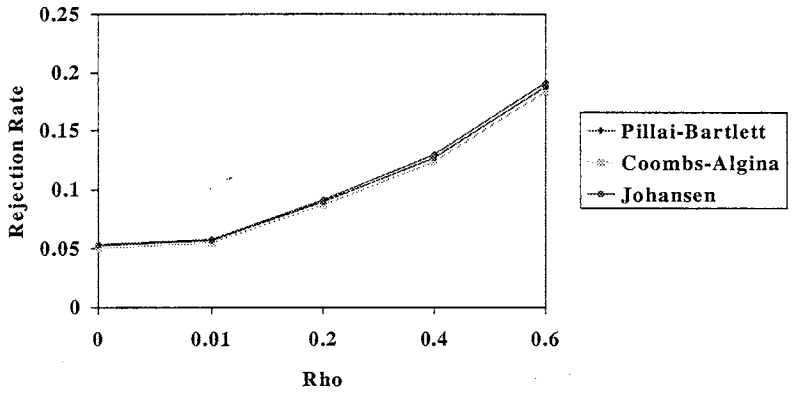


Figure 23
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=3 and Number of Subjects per Subgroup=6

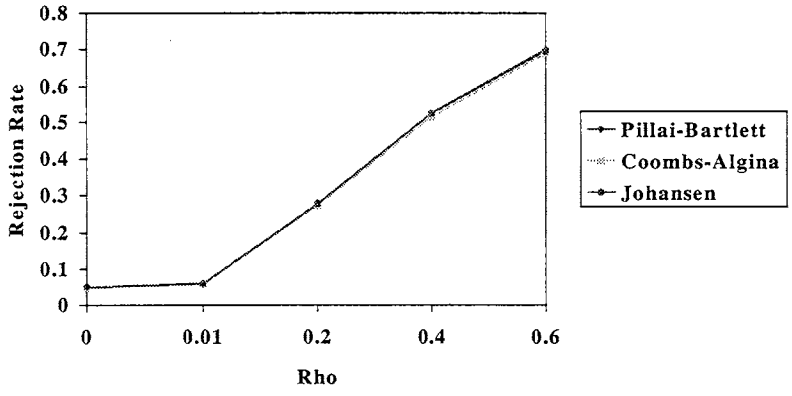


Figure 24
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Subgroup=2

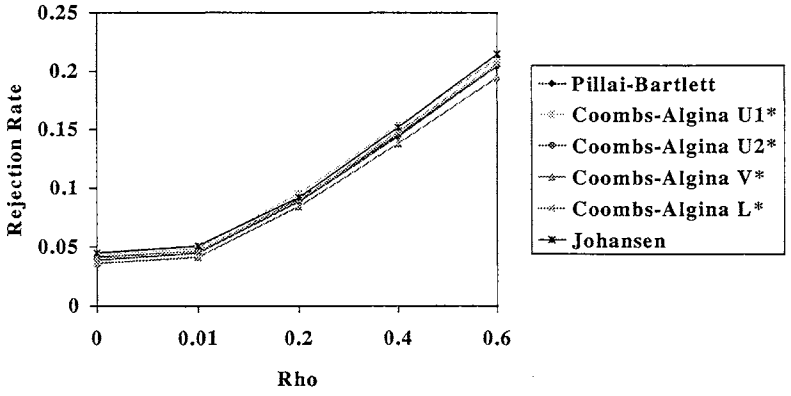


Figure 25
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Subgroup=6

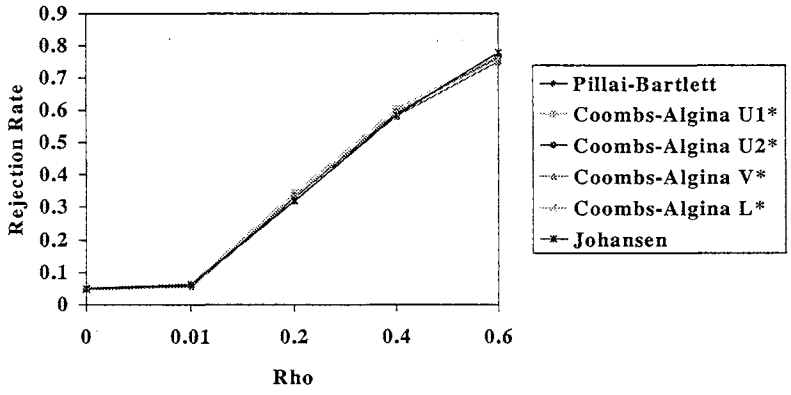


Figure 26
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=3, Number of Dependent Variables=3 and Number of Subjects per Subgroup=2

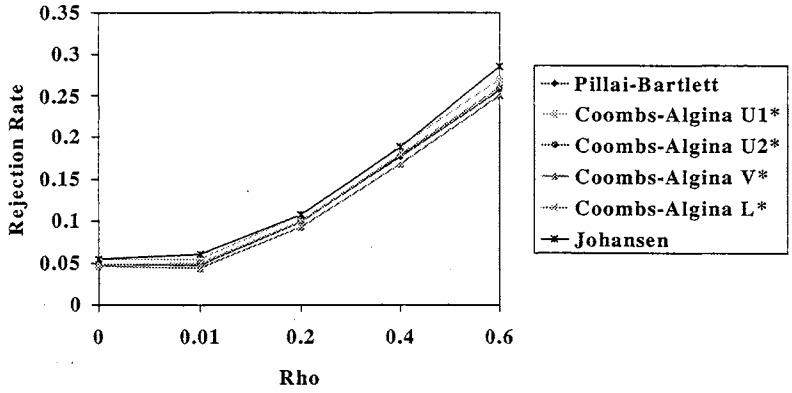
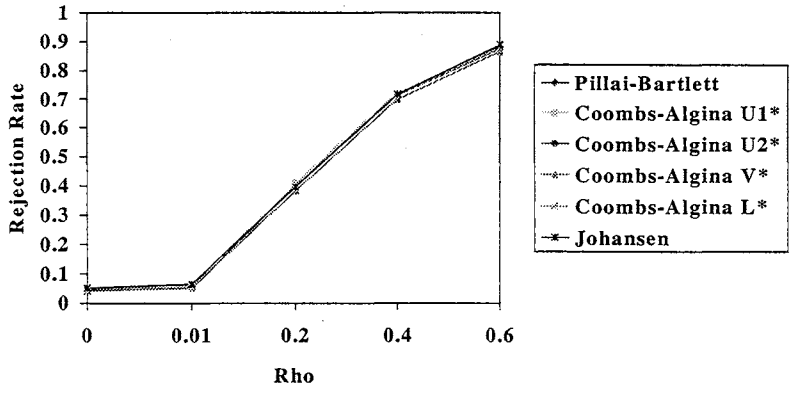


Figure 27
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=3, Number of Dependent Variables=3 and Number of Subjects per Subgroup=6



A fourth significant five-way interaction was Test x Nsubpgr x Rho x Effect Size x G. Figures 28 through 34 show that (a) as Rho increases, so does the rejection rate; (b) as Effect Size increases, so does rejection rate; (c) for positive

Effect Sizes (i.e. a real difference between groups), as the number of groups increases, the rejection rate decreases; (d) the differences in rejection rates between tests are minimal. In each of these figures the number of subjects per subgroup was held constant at Nspsub=2, the number of dependent variables was fixed at P=2, and the noncentrality parameter was fixed to simulate concentrated noncentrality.

Figure 28
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

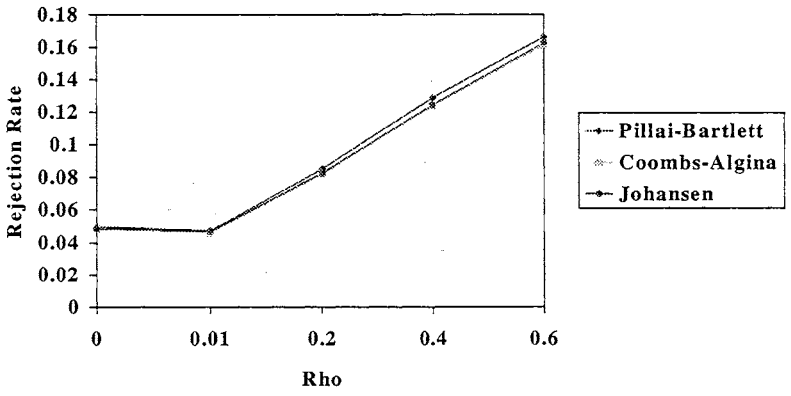


Figure 29
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=0, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24

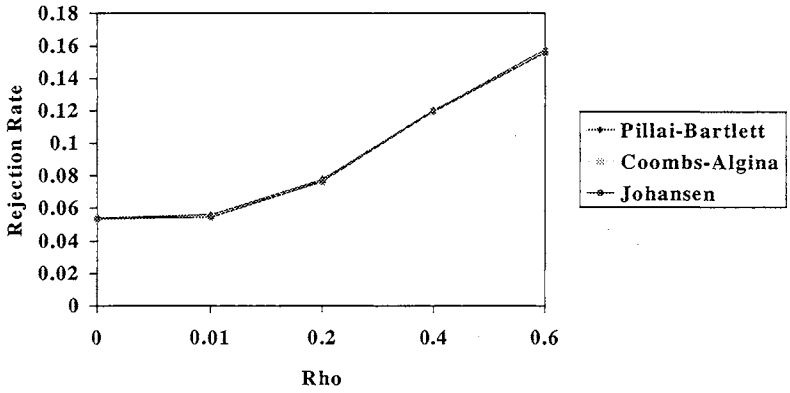


Figure 30
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=40, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

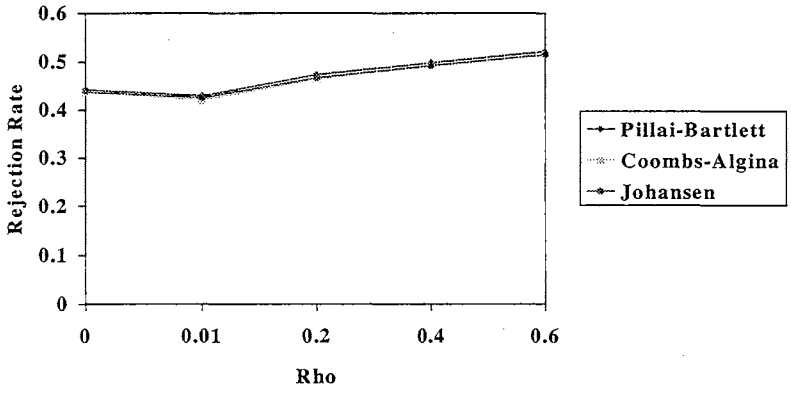


Figure 31
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=40, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24

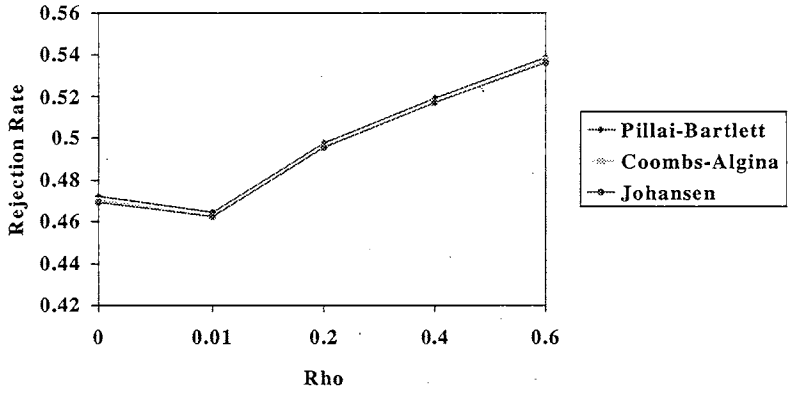


Figure 32
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=90, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

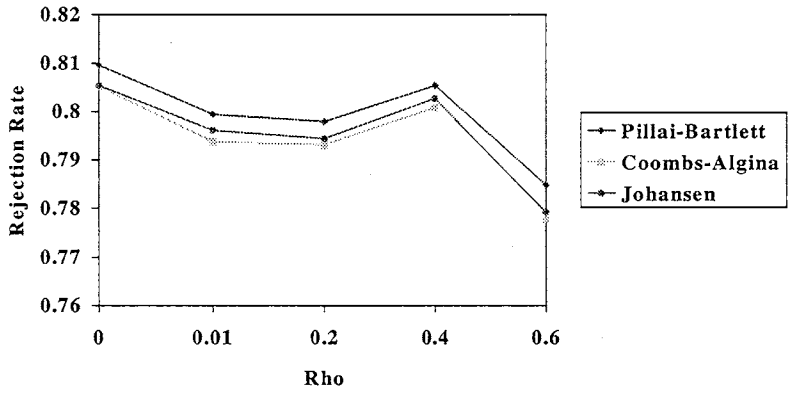


Figure 33
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=90, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24

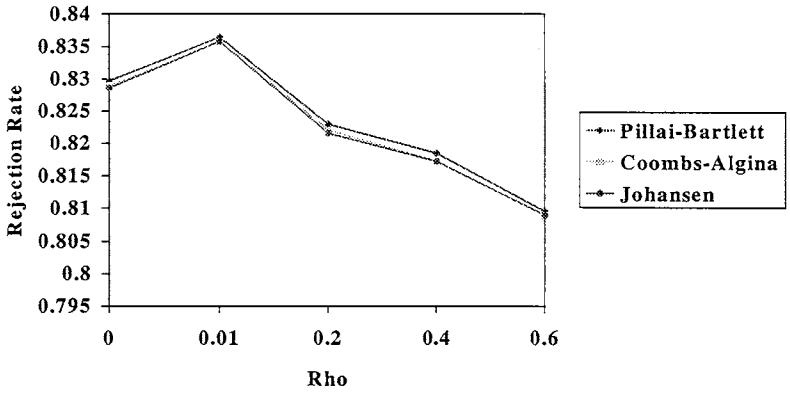
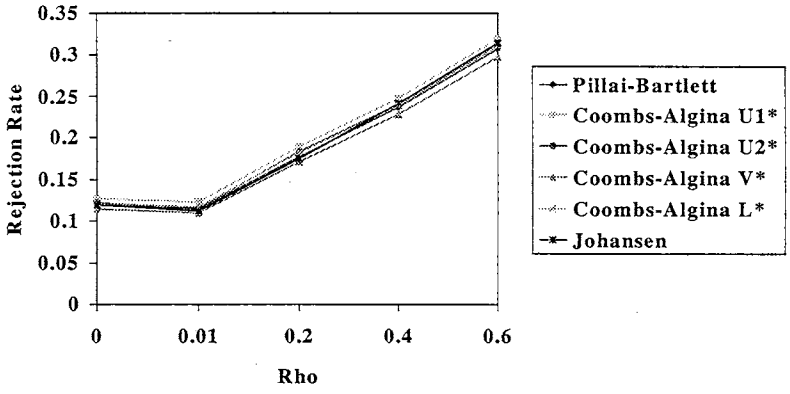


Figure 34
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=40, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12



A fifth significant five-way interaction was Test x G x NCP x Effect Size x Rho. Figures 35 through 38 show that (a) as Rho increases, so does the rejection rate; (b) the differences in rejection rates between tests are minimal; (c)

as Effect Size increases, so does the rejection rate; (d) for positive Effect Sizes, as the number of groups increases, the rejection rate decreases; (e) there is little difference between the concentrated and diffuse noncentrality conditions. In each of these figures the number of dependent variables was held constant at P=2, the number of subjects per treatment group was fixed at Nsubpgr=12, and the number of subjects per subgroup was fixed at Nspsub=2.

Figure 35
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests
With Effect Size=40, Concentrated Noncentrality, Number of Groups=2,
and Number of Dependent Variables=2

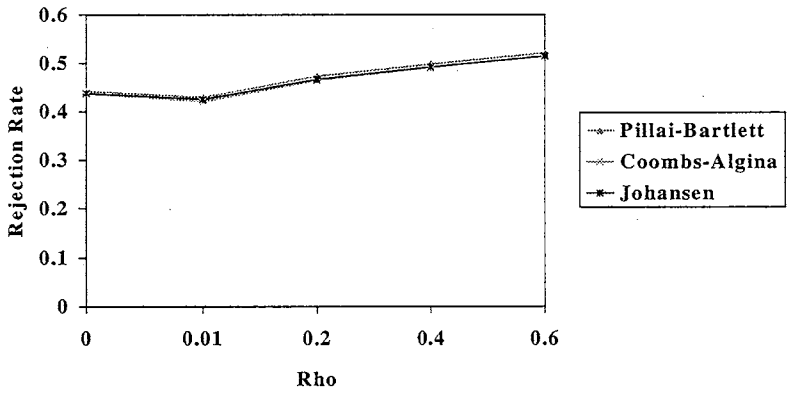


Figure 36
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=40, Diffuse Noncentrality, Number of Groups=2, and Number of Dependent Variables=2

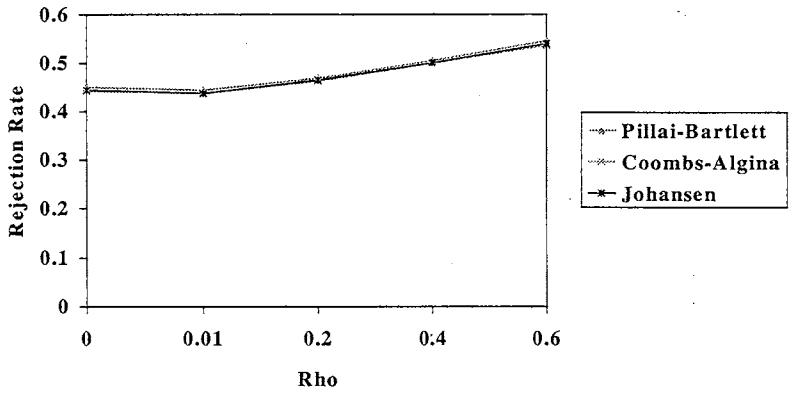


Figure 37
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=40, Concentrated Noncentrality, Number of Groups=3, and Number of Dependent Variables=2

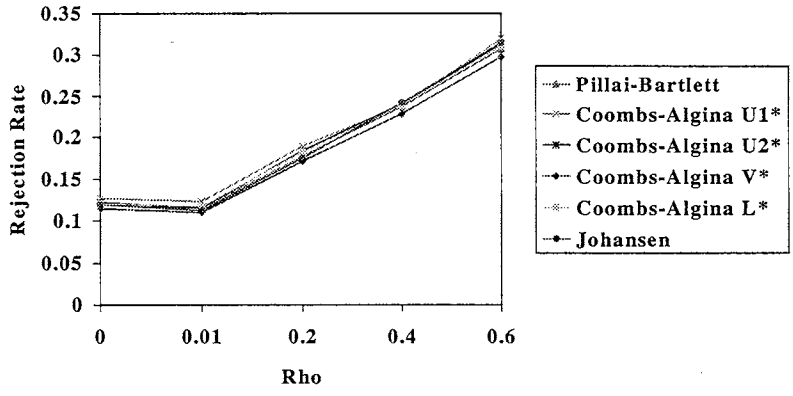
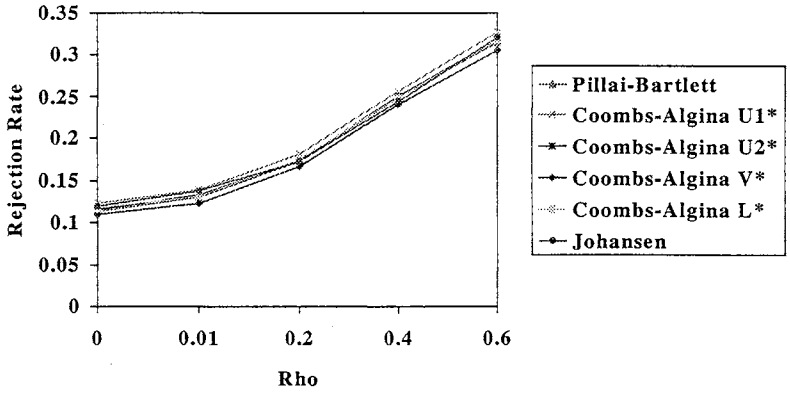


Figure 38
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Effect Size=40, Diffuse Noncentrality, Number of Groups=3, and Number of Dependent Variables=2



A sixth significant five-way interaction was Test x G x NCP x Nsubpgr x Effect Size. Figures 39 through 46 show that (a) as the number of groups increases, increasing the Effect Size has a diminishing effect on increasing the rejection rates; (b) under conditions of concentrated noncentrality, rejection rates were slightly less than those under diffuse noncentrality; (c) differences in rejection rates between tests were minimal; (d) as Effect Size increased, so did rejection rates; (e) as the size of the treatment groups increased, so did rejection rates, and this effect was more pronounced for two groups than for three groups.

Figure 39
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Concentrated Noncentrality, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

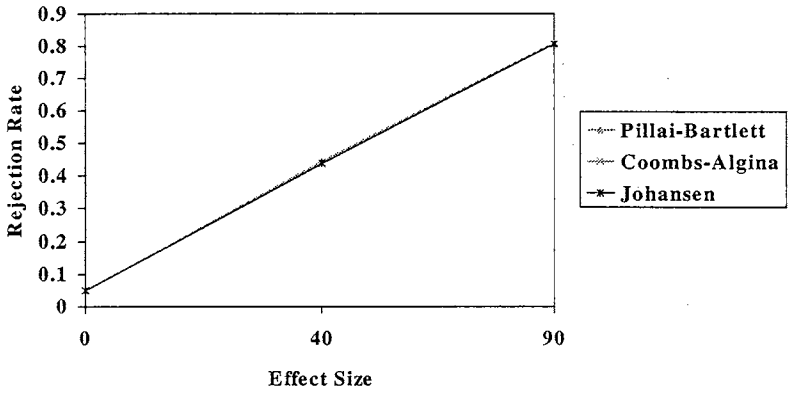


Figure 40
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Concentrated Noncentrality, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24

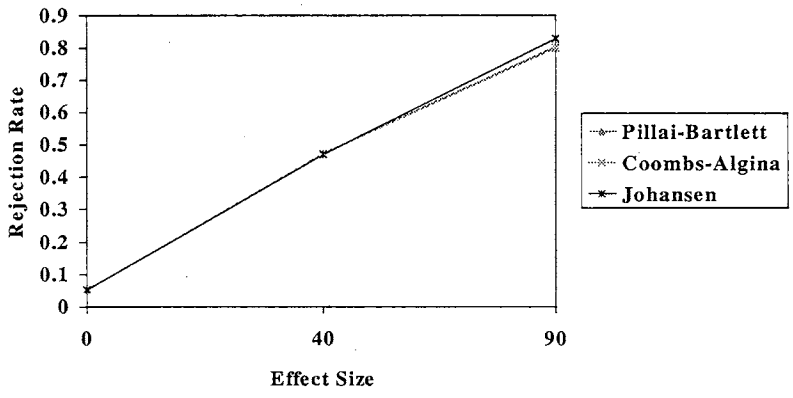


Figure 41
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Diffuse Noncentrality, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

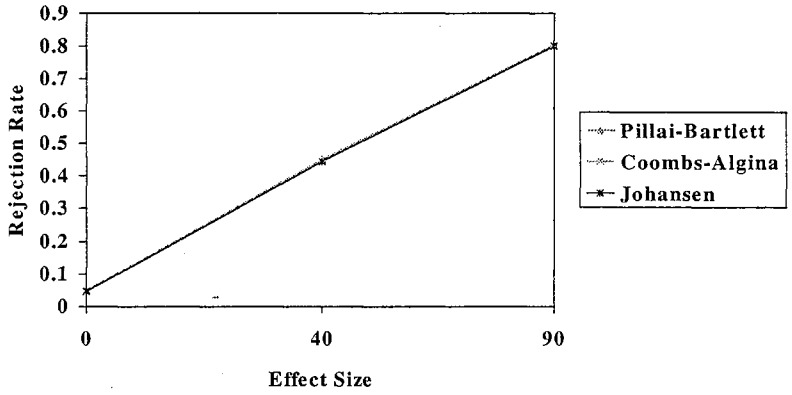


Figure 42
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Diffuse Noncentrality, Number of Groups=2, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24

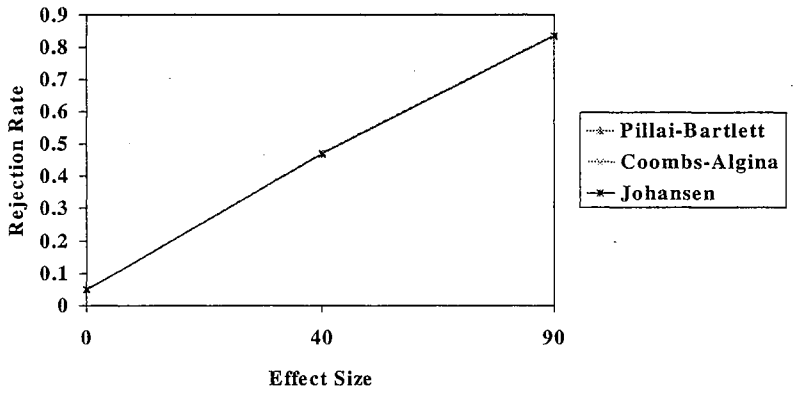


Figure 43
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Concentrated Noncentrality, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

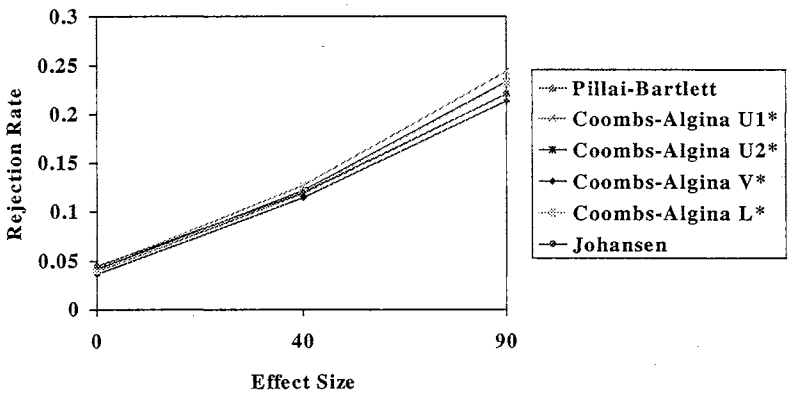


Figure 44
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Concentrated Noncentrality, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24

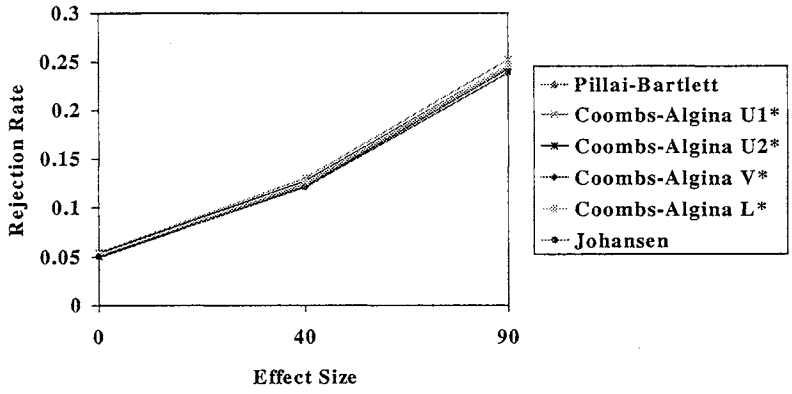


Figure 45
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Diffuse Noncentrality, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=12

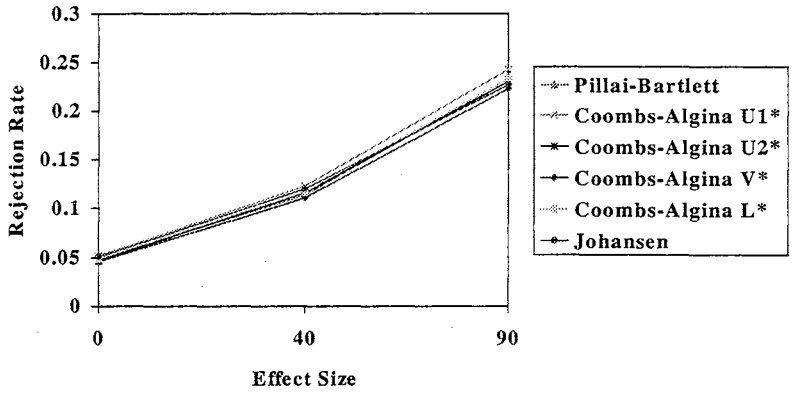
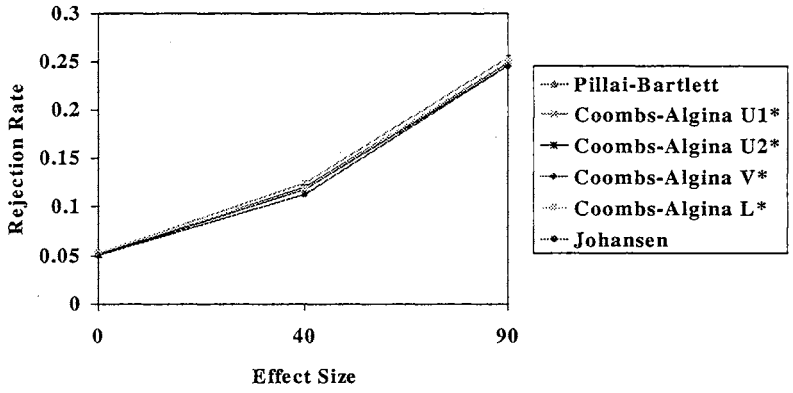


Figure 46
Comparison of Pillai-Bartlett, Coombs-Algina and Johansen Tests With Diffuse Noncentrality, Number of Groups=3, Number of Dependent Variables=2 and Number of Subjects per Treatment Group=24



The seventh significant five-way interaction was G x P x Nsubpgr x Nspsub x Rho. Figures 47 through 54 show that (a) as the number of groups increased, so did the rejection rates; (b) as the number of dependent variables

increased, so did the rejection rates; (c) as Rho increased, so did rejection rates; (d) as the size of the subgroups increased, so did rejection rates; and (e) as the size of the treatment groups increased, rejection rates went down slightly. In each of these figures, the Effect Size was fixed at zero, and the rejection rates used were averaged over the results for the six statistical tests.

Figure 47
Comparison of Rejection Rates for Values of Nspsub and Rho
With Number of Groups=2, Number of Dependent Variables=2 and
Number of Subjects per Treatment Group=12

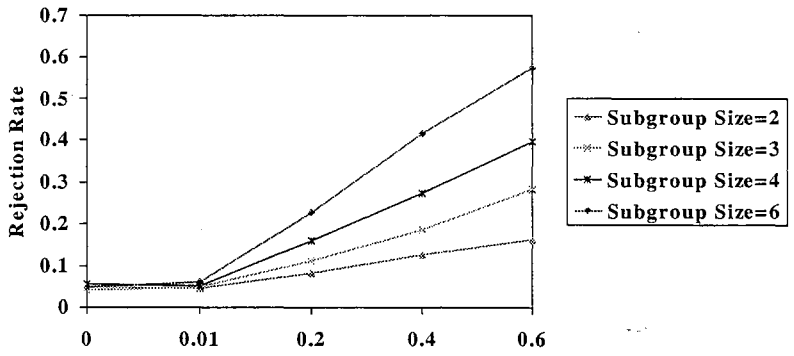


Figure 48
Comparison of Rejection Rates for Values of Nspsub and Rho
With Number of Groups=2, Number of Dependent Variables=2 and
Number of Subjects per Treatment Group=24

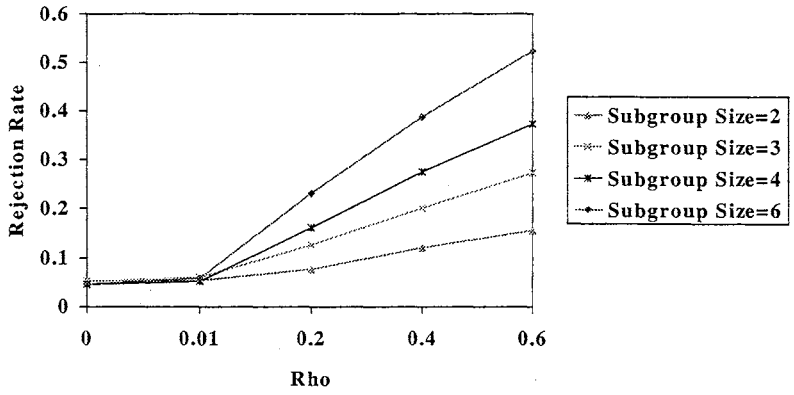


Figure 49
Comparison of Rejection Rates for Values of Nspsub and Rho
With Number of Groups=2, Number of Dependent Variables=3 and
Number of Subjects per Treatment Group=12

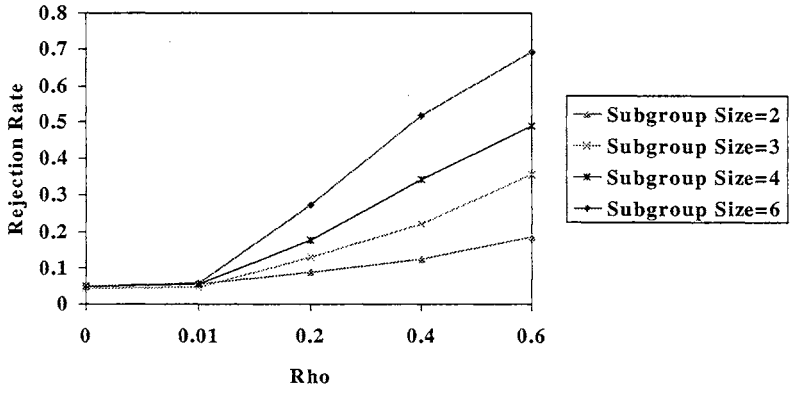


Figure 50
Comparison of Rejection Rates for Values of Nspsub and Rho
With Number of Groups=2, Number of Dependent Variables=3 and
Number of Subjects per Treatment Group=24

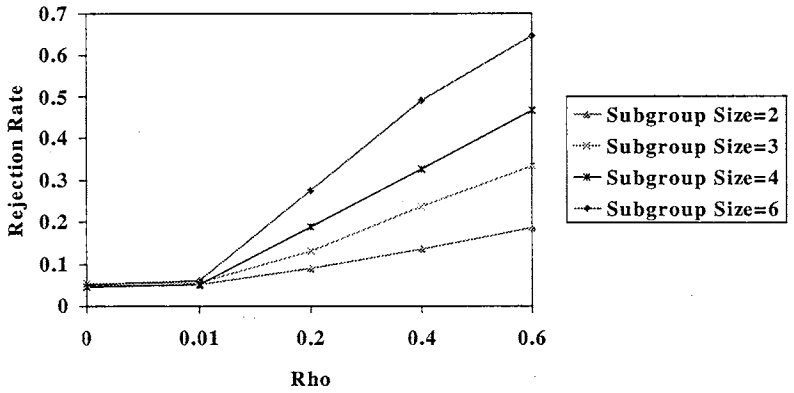


Figure 51
Comparison of Rejection Rates for Values of Nspsub and Rho
With Number of Groups=3, Number of Dependent Variables=2 and
Number of Subjects per Treatment Group=12

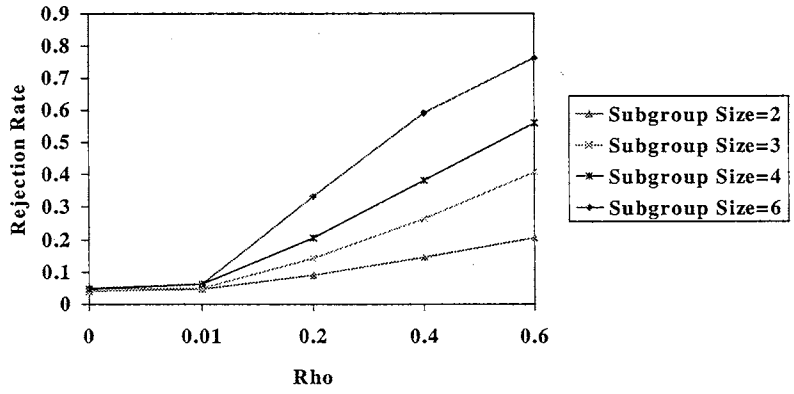


Figure 52
Comparison of Rejection Rates for Values of Nspsub and Rho
With Number of Groups=3, Number of Dependent Variables=2 and
Number of Subjects per Treatment Group=24

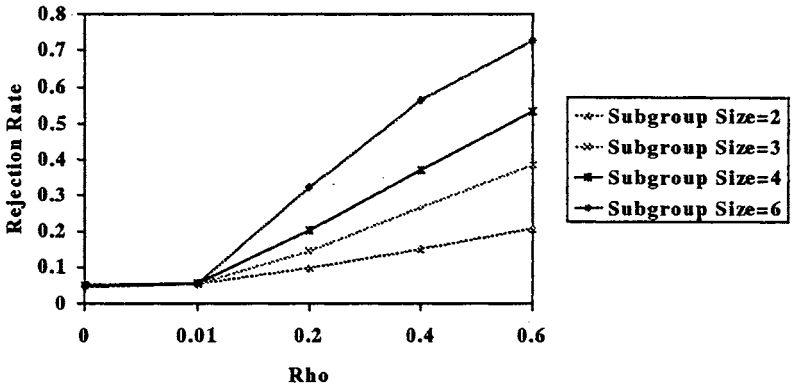


Figure 53
Comparison of Rejection Rates for Values of Nspsub and Rho
With Number of Groups=3, Number of Dependent Variables=3 and
Number of Subjects per Treatment Group=12

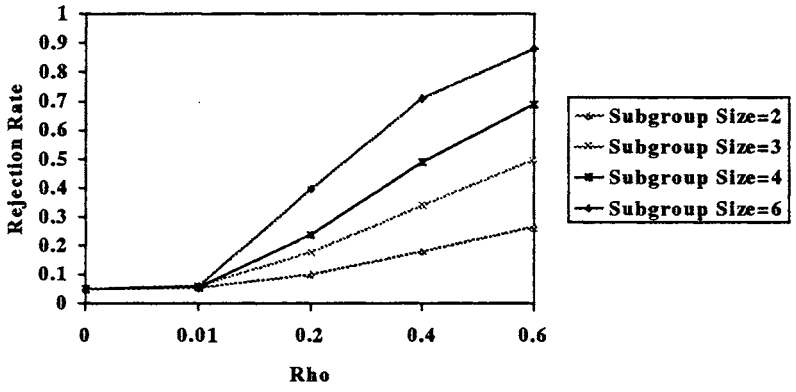
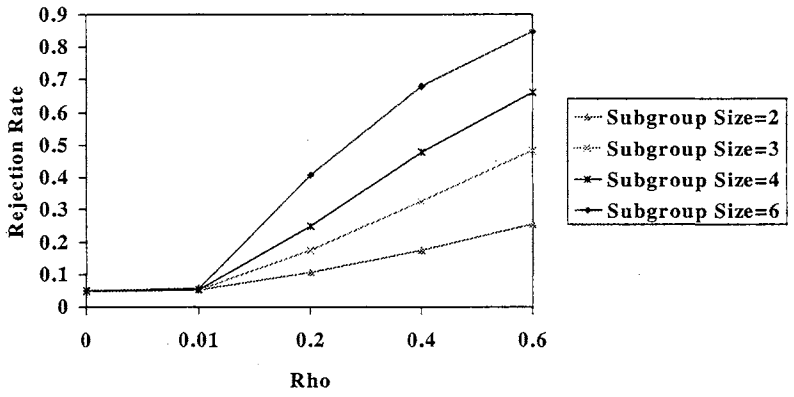


Figure 54
Comparison of Rejection Rates for Values of Nspsub and Rho
With Number of Groups=3, Number of Dependent Variables=3 and
Number of Subjects per Treatment Group=24



All of the remaining significant effects are subsumed under these seven significant five-way interactions (which together involve all eight independent variables in the model). Most of these interactions appear to be ordinal in nature, with the exception of data comparing the performance of each individual test. Rejection rates for different tests were often disordinal, but as noted in the discussions of each five-way interaction, the differences in rejection rates between the tests were usually quite small in practical terms (under 1%).

One of the more interesting results emerging from examination of the significant five-way interactions involves the relationship between Rho, the number of subjects per subgroup, the number of subjects per treatment group and the number of groups. Figure 11 showed that the intraclass correlation did not always increase with increases in each of these variables. In particular, for an Effect Size of zero, the intraclass correlation goes down slightly as the

number of subjects per treatment group goes from 12 to 24 and as the value of Rho increases. Table 3 examines this relationship in more detail.

Table 3. Actual Intraclass Correlation for Different Values of Rho, Nspsub, G and Nsubpgr

Rho	Nspsub	G=2	G=2	G=3	G=3
		Nsubpgr =12	Nsubpgr =24	Nsubpgr =12	Nsubpgr =24
.01	2	-.0044	.0017	.0011	.0018
	3	-.0029	.0017	.0021	.0017
	4	-.0014	.0017	.0027	.0017
	6	.0001	.0017	.0032	.0017
.2	2	.0433	.0403	.047	.043
	3	.0571	.0465	.0643	.0502
	4	.0709	.054	.08	.057
	6	.0965	.0699	.1116	.0733
.4	2	.0897	.0799	.0972	.0843
	3	.117	.096	.1293	.0972
	4	.1472	.1108	.1607	.1164
	6	.1977	.1376	.2198	.146
.6	2	.139	.1216	.1479	.1227
	3	.1811	.1444	.197	.1492
	4	.2201	.1653	.2419	.1723
	6	.2954	.2063	.3299	.2192

Table 3 shows that, indeed, for constant values of Rho, G and Nspsub (and with no differences between population means), the intraclass correlation does decrease as we move from treatment groups of size 12 to treatment groups of size 24. In all other respects, the intraclass correlation values increase with increases in Rho, G and Nspsub. Note that in Table 3, the intraclass correlation values are averaged over the dependent variables.

Chapter 5

Discussion

The results of this study offer some insights into the behavior of six multivariate tests under a variety of conditions primarily simulating violation of the assumption of independence of observations. Recall that the research questions posed in Chapter 1 asked (a) whether rejection rates differ as a function of the statistical test, the number of groups, the number of dependent variables, the group size, the subgroup size, the intraclass correlation (here we used a close proxy, the value we have termed Rho), the effect size, and the type of noncentrality; and (b) under what conditions each test maintains adequate control of Type I error rate and power. Several conclusions can be drawn from the results presented in Chapter 4.

Conclusion 1. Though differences between the six statistical tests studied were statistically significant in our analysis of the results, the differences were slight in practical terms, resulting in rejection rates within a range of one percentage point in nearly all cases. Thus there would be no particular advantage in recommending the use of any one test over the others under the conditions studied here.

Conclusion 2. As hypothesized, increases in the intraclass correlation dramatically increase rejection rates. Even relatively small values of the intraclass correlation (e.g. ICC=0.10) result in unacceptably high Type I error rates for conditions in which the population group means are equal.

Conclusion 3. By focusing on dependence of observations due to subgroups within treatment groups, and because dependence was operationalized in this study by varying the subgroup size, treatment group size, and degree of dependence within subgroups, we are able to see a more detailed pattern than would have been possible by simply looking at the behavior of the statistical tests under increases in the intraclass correlation. In particular, for equal population group means (i.e., no effect size), (a) increases in the size of the subgroups result in substantial increases in Type I error rates for fixed values of Rho; (b) increases in the degree of dependence within subgroups (Rho) result in substantial increases in Type I error rates; and (c) increases in the size of the treatment groups result in decreases in the Type I error rates for fixed values of Rho. The presence of significant interactions involving these variables (and others) was primarily due to differences in rates of increase in these variables, not due to disordinal interactions.

Conclusion 4. When population group means are equal, and all other variables are held constant, increases in the number of treatment groups and increases in the number of dependent variables increase the Type I error rate.

Conclusion 5. When population group means are different (i.e. a positive effect size), we find that while increases in effect size (holding other variables constant) increase rejection rates, for fixed effect sizes the rejection rates decrease when the number of groups increases. This is true for both concentrated and diffuse noncentrality conditions.

Conclusion 6. When population group means are different, rejection rates are slightly higher under conditions of diffuse noncentrality than under conditions of concentrated noncentrality. Under the group of conditions studied here, the differences between these two conditions is slight.

Limitations of this Study.

One of the primary goals of this study has been to simulate conditions of dependence of observations as they might be found in educational settings. While formalized conditions of dependence (such as cooperative group learning) may be corrected for by changing the unit of analysis, many informal situations may exist in classrooms which cannot be corrected by design. We have hypothesized that for intact classrooms, informal study groups would be the primary source of such dependence. This study has attempted to contribute to knowledge of how such groups affect rejection rates for statistical tests. However, this study does not examine the degree to which dependence actually exists in classroom settings for different types of outcome measures.

This study is also limited by the particular values chosen for Rho (0, 0.01, 0.2, 0.4, and 0.6), Nsubpgr (12 and 24), Nspsub (2, 3, 4, and 6), G (2 and 3), P (2 and 3), effect size (0, 40 and 90), and type of noncentrality (concentrated and diffuse), as well as the types of statistical tests compared. In particular, the James tests were omitted from this study, primarily due to their computational intensity.

Finally, this study is limited by the particular types of violations simulated. Heterogeneity of variance was not simulated here, nor were violations of normality.

Suggestions for Further Research.

The SAS computer program developed for this study (contained in Appendix B) was specifically designed to be versatile in accommodating many different levels of conditions without extensive coding changes. For example, any number of groups or dependent variables may be specified simply by changing a parameter. Therefore, from a technical standpoint, it would be a simple matter for other researchers to copy this program and alter the parameters to investigate a wide variety of conditions. This versatility comes at the cost of increased complexity in the computer code itself, so any researcher wishing to adapt the program for his or her own use would have a fairly steep learning curve to face before beginning a simulation.

One other technical problem encountered in this study was the length of time it takes for the program to run. Even with the limited number of conditions simulated here, the program took nearly three days to complete on the fastest personal computer available to the author. However, for a researcher willing to invest the time to understand the program, and able to spare large chunks of continuous time on a top-of-the-line personal computer, the approach developed here would be much more fruitful than previous designs which required extensive recoding for each new condition.

Besides extending the conditions studied here to more values of each variable (e.g. allowing the number of dependent variables to vary from 2 to 10), there remain a variety of conditions that have not yet been fully explored in Monte Carlo studies of the Coombs-Algina and Johansen tests. Given the promising results that have been obtained so far (Coombs, Algina & Oltman, 1996), it is important that these conditions be examined. These alternative tests will never be widely used by researchers until they are available as options on standard software packages, and one of the prerequisites for their adoption by software packages will be a thorough examination of their performance under a wide variety of conditions. It is hoped that this study will contribute to knowledge of the performance of these statistical tests, and that the computer program developed here can be used as the basis for extensive further studies.

References

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. Journal of Educational Statistics, 19, 91-101.

Algina, J. & Oshima, T. C. (1990). Robustness of the independent sample Hotelling's T^2 to variance-covariance heteroscedasticity when sample sizes are unequal or in small ratios. Psychological Bulletin, 108, 308-313.

Algina, J., Oshima, T. C., & Lin, W.-Y. (1994). Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. Journal of Educational and Behavioral Statistics, 19, 275-291.

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

Algina, J., & Tang, K. L. (1988). Type I error rates for Yao's and James's tests of equality of mean vectors under variance-covariance heteroscedasticity. Journal of Educational Statistics, 13, 281-290.

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. New York: Wiley.

Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. Biometrika, 35, 88-96.

Bartlett, M. S. (1939). A note on tests of significance in multivariate analysis. Proceedings of the Cambridge Philosophical Society, 35, 180-185.

Becker, B. (1987). Applying tests of combined significance in meta-analysis. Psychological Bulletin, 102, 164-171.

Behrens, W. U. (1929). Ein Betrag zur Fehlerberechnung bei wenigen Beobachtungen. Landwitrsch Jahrbucher, 68, 607-837.

Bock, R. D. (1975). Multivariate Statistical Methods in Behavioral Research. New York: McGraw-Hill.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. Psychological Bulletin, 57(1), 49-64.

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

Box, G. E. P. (1954b). Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics, 25, 484-498.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. Technometrics, 16(1), 129-132.

Christensen, W. F., & Rencher, A. C. (1995, August). A comparison of Type I error rates and power levels for seven solutions to the multivariate Behrsne-Fisher problem. Paper presented at the meeting of the American Statistical Association, Orlando, FL.

Clinch, J. J., & Keselman, H. J. (1982). Parametric alternatives to the analysis of variance. Journal of Educational Statistics, 7, 207-214.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Coombs, W. T. & Algina, J. (1996a). New test statistics for MANOVA/descriptive discriminant analysis. Educational and Psychological Measurement, 56(3), 382-402.

Coombs, W. T. & Algina, J. (1996b). On sample size requirements for Johansen's Test. Journal of Educational and Behavioral Statistics, 21(2), 169-178.

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. Review of Educational Research, 66(2), 137-179.

Dijkstra, J. B., & Werter, S. P. J. (1981). Testing the equality for several means when the population variances are unequal. Communication in Statistics: Simulation and Computation, B10, 557-569.

Elliott, R. S., & Barcikowski, R. S. (1994). Investigation of power using *F* approximations for the Hotelling-Lawley trace and Pillai's trace. Mid-Western Educational Researcher, 7, 2-6.

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

Fisher, R. A. (1935). The fiducial argument in statistical inference. Annals of Eugenics, 6, 391-398.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance. Review of Educational Research, 42, 237-288.

Gnanadesikan, R. (1977). Methods for Statistical Analysis of Multivariate Observations. New York: Wiley.

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

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. Journal of Educational Statistics, 17, 315-339.

Havlicek, L. L., & Peterson, N. L. (1974). Robustness of the t test: A guide for researchers on effect of violations of assumptions. Psychological Reports, 34, 1095-1114.

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 bivariate T^2 and homoscedasticity criterion M under unequal variance and leptokurtosis. Journal of the American Statistical Association, 58, 1048-1053.

Horsnell, G. (1953). The effect of unequal group variances on the *F*-test for the homogeneity of group means. Biometrika, 40, 128-136.

Hotelling, H. (1931). The generalization of Student's ratio. Annals of Mathematical Statistics, 2, 360-378.

Hotelling, H. (1951). A generalized *T* test and measure of multivariate dispersion. In J. Neyman (Ed.), Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability (pp. 23-41). Berkeley: University of California Press.

Hsu, P. L. (1938a). Contributions to the theory of 'Student's' *t*-test as applied to the problem of two samples. Statistical Research Memoirs, 2, 1-24.

Ito, K. (1969). On the effect of heteroscedasticity and non-normality upon some multivariate test procedures. In P. R. Krishnaiah (Ed.), Multivariate Analysis - II (pp. 87-120). New York: Academic Press.

Ito, K., & Schull, W. J. (1964). On the robustness of the T^2 test in multivariate analysis of variance when variance-covariance matrices are not equal. Biometrika, 38, 324-329.

James, G. S. (1951). The comparison of several groups of observations when the ratios of population variances are unknown. Biometrika, 38, 324-329.

James, G. S. (1954). Tests of linear hypotheses in univariate and multivariate analysis when the ratios 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.

Kenny, D. A., 7 Judd, C. M (1986). Consequences of violating the independence assumption in analysis of variance. Psychological Bulletin, 99(3), 422-431.

Kim, S. (1992). A practical solution to the multivariate Behrens-Fisher problem. Biometrika, 79, 171-176.

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variables. Journal of Experimental Education, 43, 61-69.

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

Lahey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlation: There's more there than meets the eye. Psychological Bulletin, 93, 586-595.

Lawley, D. N. (1938). A generalization of Fisher's z-test. Biometrika, 30, 180-187.

Lee, A. F. S., & Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. Journal of the American Statistical Association, 70, 933-941.

Lin, W.-Y. (1991). Robustness of two multivariate tests to variance-covariance heteroscedasticity and nonnormality when total-sample-size-to-variable ratio is small. (Doctoral dissertation, University of Florida, 1991). Dissertation Abstracts International, 52, 2899A.

Mardia, K. V. (1971). The effect of nonnormality on some multivariate tests and robustness to nonnormality in the linear model. Biometrika, 58, 105-121.

Mardia, K. V. (1975). Assessment of multinormality and the robustness of Hotelling's T^2 test. Applied Statistics, 24, 163-171.

Myers, J. L., DiCecco, J. V., & Lorch, R. F., Jr. (1981). Group dynamics and individual performances: Pseudogroup and quasi-*F* analyses. Journal of Personality and Social Psychology, 40, 86-98.

Nel, D. G., & van der Merwe, C. A. (1986). A solution to the multivariate Behrens-Fisher problem. Communications in Statistics: Theory and Methods, 15(12), 3475-3487.

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

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 83(4), 579-586.

Olson, C. L. (1979). Practical considerations in choosing a MANOVA test statistic: A rejoinder to Stevens. Psychological Bulletin, 86(6), 1350-1352.

Oltman, D. O. (1996). A comparison of the Type I error rates and power levels of selected multivariate analysis of variance procedures. (Doctoral dissertation, Oklahoma State University, 1996). Dissertation Abstracts International, ??

Oshima, T. C., & Algina, J. (1992a). A SAS program for testing the hypothesis of the equal means under heteroscedasticity: James's second-order test. Educational and Psychological Measurement, 52, 117-118.

Oshima, T. C., & Algina, J. (1992b). Type I error rates for James's second-order test and Wilcox's Hm test under heteroscedasticity and nonnormality. British Journal of Mathematical and Statistical Psychology, 45, 255-263.

Pavur, R., & Nath, R. (1984). Exact *F* tests in an ANOVA procedure for dependent observations. Multivariate Behavioral Research, 19, 408-420.

Pedhazur, E. J. (1982). Multiple Regression in Behavioral Research: Explanation and Prediction (3rd ed.). New York: Holt, Rinehart, and Winston.

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. Annals of Mathematical Statistics, 26, 117-121.

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

Pillai, K. C. S., & Sudjana (1975). Exact robustness studies of tests of two multivariate hypotheses based on four criteria and their distribution problems under violations. The Annals of Statistics, 3, 617-636.

Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. Journal of the American Statistical Association, 59, 665-680.

Ramsey, P. H. (1980). Exact Type I error rates for robustness of Student's *t* test with unequal variances. Journal of Educational Statistics, 5, 337-349.

Rogan, J. C., & Keselman, H. J. (1977). Is then ANOVA *F*-test robust to variance heterogeneity when sample sizes are equal?: An investigation via coefficient of variation. American Educational Research Journal, 14, 493-498.

Rubin, S. R. (1982). The use of weighted contrasts in analysis of models with heterogeneity of variance. American Statistical Association: Proceedings of the Business and Economic Statistics Section, 347-352.

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

Scheffe, H. (1959). The Analysis of Variance. New York: Wiley.

Shavelson, R. J. (1988). Statistical Reasoning for the Behavioral Sciences (2nd ed.). Boston, MA: Allyn and Bacon.

Stevens, J. P. (1979). Comment to Olson: Choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 86(2), 355-360.

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

Stevens, J. P. (1996). Applied Multivariate Statistics for the Social Sciences (3rd ed.). Mahwah, NJ: Erlbaum.

Subrahmaniam, K., & Subrahmaniam, K. (1973). On the multivariate Behrens-Fisher problem. Biometrika, 60, 107-111.

Tang, K. L., & Algina, J. (1993). Performance of four multivariate tests under variance-covariance heteroscedasticity. Multivariate Behavioral Research, 28(4), 391-405.

Tomarken, A., & Serlin, R. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. Psychological Bulletin, 99, 90-99.

Wang, Y. Y. (1971). Probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem. Journal of the American Statistical Association, 66, 605-608.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. Biometrika, 29, 350-362.

Welch, B. L. (1947). The generalization of 'Students' problem when several different population variances are involved. Biometrika, 34, 23-35.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. Biometrika, 38, 330-336.

Wilcox, R. R. (1988). A new alternative to the ANOVA *F* and new results on James' second-order method. British Journal of Mathematical and Statistical Psychology, 41, 109-117.

Wilcox, R. R. (1990). Comparing the means of two independent groups. Biometric Journal, 32, 771-780.

Wilcox, R. R. (1992). Comparing one-step *m*-estimators of location corresponding to two independent groups. Psychometrika, 57(1), 141-154.

Wilcox, R. R. (1993a). Comparing one-step m-estimators of location when there are more than two groups. Psychometrika, 58(1), 71-78.

Wilcox, R. R. (1993b). Heteroscedastic ANOVA using trimmed means versus means. Unpublished manuscript.

Wilcox, R. R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and

designing simulation studies. British Journal of Mathematical and Statistical Psychology, 48, 99-114.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F^* statistics. Communications in Statistics: Simulation and Computation, 15(4), 933-944.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. Biometrika, 24, 471-494.

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

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

Appendix A: Derivation of Formulas Creating Dependence of Observations

Let subscript *i* represent a particular treatment group, subscript *j* represent a particular subgroup, and subscript *k* represent a particular observation within the subgroup. We wish to find constants C_1 and C_2 such that the error term ε_{ijk} for our observation is given by

$$\varepsilon_{ijk} = C_1 * F_{ij} + C_2 * Z_{ik}$$

where one F_{ij} is randomly sampled for each subgroup, one Z_{ik} is randomly sampled for each observation, both F_{ij} and Z_{ik} are sampled from distributions with $\mu=0$ and SD=1, and such that

(1) $SD(\varepsilon_{ijk})=1$

(2) $Corr(\varepsilon_{ijk}, \varepsilon_{ijl})=\rho$ for *k*, *l* in the same subgroup

(3) $E(\varepsilon_{ijk})=0$, and

(4) $Corr(\varepsilon_{ijk}, \varepsilon_{iml})=0$ for *j*, *m* different subgroups.

To do this we let

$$\rho = \frac{b^2}{b+1}$$

$$C_1 = \frac{b}{\sqrt{b^2+1}}$$

$$C_2 = \frac{1}{\sqrt{b^2+1}}$$

Then,

(1) $Var(\varepsilon_{ijk}) = Var(C_1*F_{ij} + C_2*Z_{ik})$

$= C_1^2Var(F_{ij}) + C_2^2Var(Z_{ik}) + 2C_1C_2Cov(F_{ij},Z_{ik})$

$$= \left(\frac{b}{\sqrt{b^2+1}}\right)^2 (1) + \left(\frac{1}{\sqrt{b^2+1}}\right)^2 (1) + 0$$

$= 1$

(2) $Corr(\varepsilon_{ijk}, \varepsilon_{ijl}) = Cov(C_1F_{ij}+C_2Z_{ik}, C_1F_{ij}+C_2Z_{il}) / (SD(\varepsilon_{ijk})*SD(\varepsilon_{ijl}))$

$=(C_1^2Cov(F_{ij},F_{ij}) + C_1C_2Cov(Z_{ik},F_{ij}) + C_1C_2Cov(F_{ij},Z_{il}) + C_2^2Cov(Z_{ik},Z_{il})) /$

1

$= C_1^2Var(F_{ij}) + 0 + 0 + 0$

$= \rho*1$

$= \rho$

(3) $E(\varepsilon_{ijk}) = E(C_1F_{ij}+C_2Z_{ik})$

$= C_1E(F_{ij}) + C_2E(Z_{ik})$

$= 0 + 0 = 0$

(4) $Corr(\varepsilon_{ijk}, \varepsilon_{iml}) = Cov(\varepsilon_{ijk}, \varepsilon_{iml}) / (SD(\varepsilon_{ijk})*SD(\varepsilon_{iml}))$

$= Cov(\varepsilon_{ijk}, \varepsilon_{iml}) / (1*1)$

$= Cov(C_1F_{ij}+C_2Z_{ik}, C_1F_{im}+C_2Z_{il})$

$= C_1C_1Cov(F_{ij},F_{im}) + C_1C_2Cov(F_{ij},Z_{il}) + C_2C_1Cov(Z_{ik},F_{im}) +$

$C_2C_2Cov(Z_{ik},Z_{il})$

$$= 0 + 0 + 0 + 0 = 0$$

Thus, if we fix a value for ρ to simulate a level of dependence, we can solve for the constant *b*, obtain C_1 and C_2, and randomly sample F_{ij} and Z_{ik} values as specified above to obtain the error term for each observation.

Appendix B: SAS Program Used for Monte Carlo Study

```
*****************************************************************;
* Reichard & Coombs  Monte Carlo Study                          *;
*                                                               *;
* This SAS program, contained mostly in PROC IML, simulates data *;
* with the following conditions:                                *;
* 1. dependence of observations (rho) = 0.01, 0.2, 0.4, and 0.6 *;
* 2. sample size per subgroup = 2, 3, and 4 (total group sample *;
*    size fixed at 24)                                          *;
* 3. number of groups = 2, 3                                    *;
* 4. number of dependent variables = 2, 3                       *;
* 5. effect size (measured by noncentrality param) = 0, 40, 90  *;
* 6. type of noncentrality = Concentrated (first dependent vble *;
*    only), Diffuse (all dependent variables)                   *;
*                                                               *;
* These conditions will be tested using the following statistical *;
* tests:  Pillai-Bartlett, Johansen, Coombs-Algina U1*, U2*, V*, *;
* and L*.  Results will be compared to see how well each test   *;
* performs.                                                     *;
*****************************************************************;

 dm 'output; clear; log; clear; ';
*****************************************************************;
* I N T R O D U C T I O N                                       *;
*                                                               *;
* We call IML, and set up most of our program logic as a macro. *;
* The macro will then be repeatedly called for each combination *;
* of conditions that we are testing.                            *;
* Variables are initialized.                                    *;
*****************************************************************;
 proc iml symspace=150;
 show space;
%macro design(rho,delta,nsubpgr,nspsub,ngrp,ndep,ncp);
Fsig=0; PBsig=0; U1sig=0; U2sig=0;
Vsig=0; Lsig=0; Jsig=0;
reps=1000;                                    ** <=== change this! ;
Intrasum = j(&ndep,1,0);
*****************************************************************;
* Begin the Loop                                                *;
*                                                               *;
* All code inside this loop is replicated a fixed number of times *;
* as set by variable "reps".                                    *;
*****************************************************************;
do loop=1 to reps;
   rho=&rho; delta=&delta; ngrp=&ngrp;
   nsubpgr=&nsubpgr; nspsub=&nspsub; ndep=&ndep ; ncp=&ncp;


   S1 = min(ndep,ngrp);                    ** use this for ncp ;
   subgpgr = nsubpgr / nspsub; * # subgps per group ;
*****************************************************************;
* Create Error Vectors                                          *;
```

```
*                                                                    *;
* This part is fairly complicated because of the dependence of      *;
* observations. Recall that each Y score (dependent variable        *;
* score) can be expressed as Y = mu + error, where mu is the        *;
* group mean.  So the error vectors contain all of the variance     *;
* within each group. In order to simulate varying degrees of        *;
* dependence, we use the number of subjects per correlated          *;
* subgroup and the number of subgroups per group to convert         *;
* our value of rho into coefficients c1 and c2 such that each       *;
* persons  score is a combination of unique random error and        *;
* a second error component which is fixed for each subgroup but     *;
* varies between subgroups.  This is done to simulate the effects   *;
* of small groups within classrooms who study/play together.        *;
* See dissertation text for an explanation of how rho relates to    *;
* c1 and c2 and why these were chosen. Essentially we want to       *;
* make sure that the expected value of the errors is zero, and      *;
* their standard deviation is one.                                  *;
**********************************************************************;
   b=sqrt(rho/(1-rho));
   c1=b/sqrt(b**2+1);
   c2=1/sqrt(b**2+1);

   * compute group size vector;
   Nvec = j(ngrp,1,0);
   do i = 1 to ngrp;
      Nvec[i,1]=nsubpgr;
   end;
   N = Nvec[+,];  * total sample size ;
 *  print Nvec ;

   * compute vectors with cumulative group sizes ;
   * these will be used in later calculations ;
   Rvecp = j(ngrp+1,1,0);   * this one has multiples of the # of dep
vbles;
   Rveck = j(ngrp+1,1,0);   * this one has multiples of the # of
subj per group ;
   do i=1 to ngrp;
      Rvecp[i+1,1] = ndep + Rvecp[i,1];
   end;
   Rvecp = 1 + Rvecp ;
   do i=1 to ngrp ;
      Rveck[i+1,1] = Nvec[i,1] + Rveck[i,1] ;
   end;
   Rveck = 1 + Rveck ;
 *  print Rvecp ;
 *  print Rveck ;

   * ;
  Evec = j(N,ndep,0);  * generate a vector with all error terms;
   do m=1 to ndep;                          * columns ;
      do k=1 to ngrp;                       * rows ;
         ff = j(Nvec[k,1],1,0);
         do i=1 to subgpgr;
            ransub=rannor(0); * random component for each subgroup ;
            do j=1 to nspsub;
               ff[(i-1)*nspsub+j,1]=ransub;
            end;
         end;
         uu = rannor(j(Nvec[k,1],1,0)); * random comp for each subj;
         ee = c1*ff + c2*uu;
```

```
         do i=1 to Nvec[k,1];
            sum=0;
            do j=1 to (k-1);
               sum = sum + Nvec[j,1];
            end;
            row=i + sum;
            Evec[row,m] = ee[i,1];
         end;
      end;
   end;
  free uu ee ff ;

*****************************************************************;
* Generate Mean Vector                                          *;
* The mean vector is determined by both the noncentrality       *;
* structure and the effect size, as measured by the noncentrality *;
* parameter.  For concentrated noncentrality, the ncp param     *;
* appears as the group mean of the 1st group, 1st vble only.  all *;
* other means are zero.  For diffuse noncentrality, the ith     *;
* group and ith vble get a mean equal to the ncp param for i=1  *;
* through S=min(ngrp,ndep).  Other means are zero.              *;
*****************************************************************;


  Uvec = j(ngrp,ndep,0);   * generate a vector with each group mean;
  if ncp='c' then do;
     ncparam = sqrt(delta/(N*ngrp*(ngrp-1)));
     Uvec[1,1]=ncparam;
  end;
  if ncp='d' then do;
     if ngrp>ndep then
        ncparam = sqrt(delta/((ndep-1)*N*(ngrp**2) + N*ngrp*(ngrp-
ndep)));
     else
        ncparam = sqrt(delta/((ngrp-1)*N*(ngrp**2)));
     do i=1 to S1;
        Uvec[i,i]=ncparam;
     end;
  end;


  * generate Y scores by adding the means and the errors ;
  Yvec = j(N,ndep,0);
  do m=1 to ndep;
     do k=1 to ngrp;
        do j=1 to Nvec[k,1];
          sum=0;
          do i=1 to (k-1);
             sum = sum + Nvec[i,1];
          end;
          row = j + sum;
          Yvec[row,m] = Evec[row,m] + Uvec[k,m];
        end;
     end;
  end;
*  Print Yvec ;
  free Evec Uvec ;

*****************************************************************;
* now we need to use the Y score vector and various subvectors *;
```

```
* to compute the sums of squares, etc. necessary to calculate  *;
* the tests we need                                             *;
****************************************************************;

   Meanvec = j(ngrp,ndep,0);
   Grand = j(1,ndep,0);
   do k=1 to ngrp;
      row1 = Rveck[k,1];
      row2 = Rveck[k+1,1] - 1;
      Ybar = Yvec[row1:row2,1:ndep];
      do j=1 to ndep;
         Meanvec[k,j] = Ybar[+,j] / Nvec[k,1] ;
         Grand[1,j] = Grand[1,j] + Ybar[+,j];
      end;
      free Ybar;
   end;
   Grand = Grand / N;
   JmeanT = Meanvec` ; * save off clean mean transpose for johansen
test ;
 *  print Meanvec Grand ;

   do m=1 to ndep;
      do k=1 to ngrp;
         do j=1 to Nvec[k,1];
            row = Rveck[k,1] + j - 1;
            Yvec[row,m] = Yvec[row,m] - Meanvec[k,m];
         end;
      end;
   end;
 *  print Yvec;

* create a matrix of SSCP matrices - one pxp matrix for each group ;
* (all nested in the large matrix SSCP) ;
   SSCP = j(ngrp*ndep,ndep,0);
   do k=1 to ngrp;
      row1 = Rveck[k,1];
      row2 = Rveck[k+1,1] - 1 ;
      Ysub = j(Nvec[k,1],ndep,0);
      Ysub[1:Nvec[k,1],1:ndep] = Yvec[row1:row2,1:ndep];
      Esub = Ysub`*Ysub;
      row1 = Rvecp[k,1];
      row2 = Rvecp[k+1,1] - 1 ;
      SSCP[row1:row2,1:ndep] = Esub;
      free Ysub Esub ;
   end;
 *  print SSCP ;

* now we need to pool the SSCP matrices ;
* this is the matrix E ;
   E = j(ndep,ndep,0);
   do m=1 to ndep;
      do k=1 to ndep;
         do j=1 to ngrp;
            row = k + (j-1)*ndep;
            E[k,m] = E[k,m] + SSCP[row,m];
         end;
      end;
   end;

*  now we need H ;
```

```
*  we subtract the grand mean from the group means ;
*  then we multiply the result by its transpose ;
*  to get the sums of squares and cross products (between) ;
   do m=1 to ndep;
      do k=1 to ngrp;
         Meanvec[k,m] = Meanvec[k,m] - Grand[1,m];
      end;
   end;

   H = j(ndep,ndep,0);
   MeanvecT = Meanvec` ;
   do m=1 to ndep ;
      do k=1 to ngrp;
         Meanvec[k,m] = Meanvec[k,m] # Nvec[k,1];
      end;
   end;
   H = MeanvecT * Meanvec ;
   T = E + H ;
 *  print H ;

* Calculate the intraclass correlation for each variable ;
* This will be used to examine how differences in rho, nspsub ;
* and nsubpgr affect the actual intraclass correlation ;

Intra = j(ndep,1,0);
do i=1 to ndep;
   Intra[i,1] = (H[i,i]/(ngrp-1) - E[i,i]/(N-ngrp))/
                (H[i,i]/(ngrp-1) + (nsubpgr-1)*E[i,i]/(N-ngrp));
end;

* print H E Intra Nvec ;

* Coombs-Algina Calculations ;
   C = Nvec # (1/N);
   C = 1 - C;
   Sampvec = Nvec - 1 ; * used for sample std dev ;
   Svec = j(ngrp*ndep,ndep,0);
   do k=1 to ngrp;
      row1 = Rvecp[k,1];
      row2 = Rvecp[k+1,1] - 1 ;
      Svec[row1:row2,1:ndep] = SSCP[row1:row2,1:ndep] /
Sampvec[k,1];
   end;
   Mvec = j(ndep,ndep,0);
   do m=1 to ndep;
      do k=1 to ndep;
         do j=1 to ngrp;
            row = k + (j-1)*ndep;
            Mvec[k,m] = Mvec[k,m] + Svec[row,m] # C[j,1];
         end;
      end;
   end;
 *  print Svec ;
   denom=0;
   do k=1 to ngrp;
      row1 = Rvecp[k,1];
      row2 = Rvecp[k+1,1] - 1;
      Ssub = Svec[row1:row2,1:ndep];
      Msub = Ssub # C[k,1];
```

```
      denom = denom + (1/Sampvec[k,1]) * (trace(Msub)*trace(Msub) +
trace(Msub**2));
   end;
   numer = trace(Mvec)*trace(Mvec) + trace (Mvec**2);
   G3 = numer / denom ;

 * calculate Johansen vectors ;
   Xsum = j(ndep,1,0);
   Wvec = j(ngrp*ndep,ndep,0);
   do k=1 to ngrp;
      row1 = Rvecp[k,1];
      row2 = Rvecp[k+1,1] - 1;
      Wvec[row1:row2,1:ndep] =
inv(Svec[row1:row2,1:ndep]/Nvec[k,1]);
      Xsum = Xsum + Wvec[row1:row2,1:ndep]*JmeanT[1:ndep,k];
   end;
   Wsum = j(ndep,ndep,0);
   do m=1 to ndep;
      do k=1 to ndep;
         do j=1 to ngrp;
            row = k + (j-1)*ndep;
            Wsum[k,m] = Wsum[k,m] + Wvec[row,m];
         end;
      end;
   end;
   Xbar = inv(Wsum)*Xsum;
 *  print Xbar Sampvec;
   Wsub = j(ndep,ndep,0);
   A=0; J=0;
   do k=1 to ngrp;
      row1 = Rvecp[k,1];
      row2 = Rvecp[k+1,1] - 1;
      Wsub = Wvec[row1:row2,1:ndep];
      J = J + (JmeanT[1:ndep,k] - Xbar)` * Wsub * (JmeanT[1:ndep,k]
- Xbar) ;
      Asub = j(ndep,ndep,0);
      Asub = I(ndep) - (inv(Wsum) * Wsub) ;
      A = A + (trace(Asub**2) + trace(Asub)*trace(Asub)) /
(2*Sampvec[k,1]) ;
   end;
   C1 = ndep*(ngrp-1) + 2*A - (6*A)/(ndep*(ngrp-1) +2);

   free Xsum Wvec Wsum Wsub Xbar Asub SvecT JmeanT ;


* calculate "parameters" *;
* these are intermediate amounts used in the * ;
* formulas for each of the tests, particularly * ;
* for the degrees of freedom * ;
   BFDF1 = ngrp - 1 ;
   MM = .5 * (abs(ndep - BFDF1)-1) ;
   NN = .5 * (G3 - ndep - 1);
   BB = ((2*NN + BFDF1)*(2*NN + ndep))/(2*(NN-1)*(2*NN + 1));
   AA = 4 + (ndep*BFDF1 + 2)/(BB-1) ;
   NNN = .5*(N - ngrp - ndep - 1);
   BBB = ((2*NNN + BFDF1)*(2*NNN + ndep))/(2*(NNN-1)*(2*NNN + 1));
   AAA = 4 + (ndep*BFDF1 + 2)/(BBB-1) ;
   G3M = G3 / (ngrp-1) ;
   S = min(ndep,ngrp-1);
```

```
* now were ready for the Pillai-Bartlett calculations ;
   PB = trace(H*inv(T));
   FPB =  PB * (2*NNN + S + 1) / ((S - PB)*(2*MM + S + 1));
   DF1PB = S*(2*MM + S + 1);
   DF2PB = S*(2*NNN + S + 1);
   ProbPB = 1 - probf(FPB,DF1PB,DF2PB);
   if ProbPB <= 0.05 then PBsig = PBsig +1 ;

* compute C-A U1 & U2 ;

 *  Temp = G3M*Mvec;
   U = trace(H*inv(G3M*Mvec));  * used for both U1 & U2 ;
   DF1U1 = S*(2*MM + S + 1);
   DF2U1 = 2*(S*NN + 1);
   FU1 = ((2*(S*NN + 1))*U)/((S*(2*MM + S + 1))*S);
   ProbU1 = 1 - probf(FU1,DF1U1,DF2U1);
   if ProbU1 <= 0.05 then U1sig = U1sig + 1;

   FU2 = ((2*NN)*AA*U)/((AA-2)*(ndep*BFDF1));
   DF1U2 = ndep*BFDF1;
   DF2U2 = AA ;
   ProbU2 = 1 - probf(FU2,DF1U2,DF2U2);
   if ProbU2 <= 0.05 then U2sig = U2sig +1 ;

* compute C-A V ;
   V = trace(H*inv(H+(G3M*Mvec)));
   FV = (2*NN + S + 1)*V / ((2*MM + S + 1)*(S-V));
   DF1V = S*(2*MM + S + 1);
   DF2V = S*(2*NN + S + 1);
   ProbV = 1 - probf(FV,DF1V,DF2V);
   if ProbV<= 0.05 then Vsig = Vsig + 1 ;

* compute C-A L ;
   L = det(G3M*Mvec)/det(H+(G3M*Mvec)) ;
   Check = ndep**2 + BFDF1**2 - 5 ;
   if Check > 0 then
      TT = sqrt(((ndep**2)*(BFDF1**2) - 4)/check) ;
   else TT = 1 ;
   RR = G3 - (ndep - BFDF1 + 1)/2 ;
   QQ = (ndep*BFDF1 - 2)/4 ;
   FL = ((1-L##(1/TT))*(RR*TT-2*QQ))/((L##(1/TT))*(ndep*BFDF1)) ;
   DF1L = ndep*BFDF1 ;
   DF2L = RR*TT - 2*QQ ;
   ProbL = 1 - probf(FL,DF1L,DF2L);
   if ProbL <= 0.05 then Lsig = Lsig + 1 ;

*  compute Johansen ;
   FJ = J / C1;
   DF1J = ndep*(ngrp-1);
   DF2J = ndep*(ngrp-1)*(ndep*(ngrp-1)+2)/(3*A);
 *  print FJ A J C1 DF1J DF2J ;
   ProbJ = 1 - probf(FJ,DF1J,DF2J);
   if ProbJ <= 0.05 then Jsig = Jsig + 1;

   Intrasum = Intrasum + Intra ;

   free Nvec Rvecp Rveck Yvec Meanvec MeanvecT Grand SSCP H E T C
Sampvec Svec Mvec Msub Intra;
```

```
end ;  ** the end of the loop ! ;

* compute actual alpha levels ;
alphaPB = PBsig/reps ;
alphaU1 = U1sig/reps ;
alphaU2 = U2sig/reps ;
alphaV  = Vsig/reps  ;
alphaL  = Lsig/reps  ;
alphaJ  = Jsig/reps  ;
Intrasum = Intrasum / reps;

* Print Intrasum alphaPB alphaU1 FPB FU1 DF1PB DF2PB ;

* output results to file ;
filename out 'c:\temp\diss1.txt' mod ;
file out ;
put rho 5.2 delta 3. nsubpgr 3. nspsub 3. ngrp 3. ndep 3. ncp $1.
    alphaPB 7.4 alphaU1 7.4 alphaU2 7.4 alphaV 7.4 alphaL 7.4
    alphaJ 7.4 ;
closefile out ;
run ;

* output intraclass corr to another file ;
filename out2 'c:\temp\diss2.txt' mod ;
file out2 ;
put rho 5.2 delta 3. nsubpgr 3. nspsub 3. ngrp 3. ndep 3. ncp $1. @;
do i=1 to ndep;
  put  (Intrasum[i,1]) 5.2 @;
end;
put;
closefile out2 ;
run ;

%mend ;

* set up matrices of design values for looping ;
des1 = {0 0.01 0.2 0.4 0.6} ;  * rho values ;
des2 = {0 40 90};              * delta values ;
des3 = {12 24};                * group size ;
des4 = {2 3 4 6};              * # subj per subgp;
des5 = {2 3};                  * # groups ;
des6 = {2 3};                  * # dep vbles ;
des7 = {'c' 'd'};              * ncp ;


do i1=1 to 5 ;                 * loop for des1 ;
  do i2=1 to 3;                * loop for des2 ;
    do i3=1 to 2;              * loop for des3 ;
      do i4=1 to 4;            * loop for des4 ;
        do i5=1 to 2;          * loop for des5 ;
          do i6=1 to 2;        * loop for des6 ;
            do i7=1 to 2;      * loop for des7 ;
              do i8=1 to 5 ;   * loop 5 times for correct error term
in ANOVA ;

%design(des1[i1],des2[i2],des3[i3],des4[i4],des5[i5],des6[i6],des7[i
7]);
              end;
            end;
```

```
          end;
        end;
      end;
    end;
  end;
end;

            /*
 %design(0.6,0,12,4,2,2,'c');
 %design(0.6,0,24,4,2,2,'c');
            */
run;

quit;  * quits proc iml ;

run ;
```

VITA

Carla Reichard

Candidate for the Degree of

Doctor of Philosophy

Dissertation: A MONTE CARLO STUDY OF SEVERAL MULTIVARIATE ANALYSIS OF VARIANCE PROCEDURES

Major Field: Applied Behavioral Studies

Biographical:

Education: Graduated from East Central High School in Tulsa, Oklahoma in May, 1976; received Bachelor of Science degree in Mathematics from the University of Oklahoma, Norman, Oklahoma in May, 1980; received Master of Arts degree in Mathematics from the University of Kansas, Lawrence, Kansas in July, 1982. Completed the requirements for the Doctor of Philosophy degree in Educational Research, Evaluation, Measurement and Statistics in May, 1999.

Experience: Employed by Oklahoma State University in the Office of Planning, Budget and Institutional Research, May 1992 to present.

Professional Memberships: Oklahoma Association for Institutional Research and Planning, MidAmerica Association for Institutional Research, Southern Association for Institutional Research, Association for Institutional Research, American Statistical Association.