A COMPARISON OF THE TYPE I ERROR RATES AND POWER LEVELS OF SELECTED MULTIVARIATE ANALYSIS OF VARIANCE PROCEDURES

By

DEBRA ANN OLTMAN

Bachelor of Arts
University of South Dakota
Vermillion, South Dakota
1971

Master of Arts
University of Nebraska
Lincoln, Nebraska
1975

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, 1996

A COMPARISON OF THE TYPE I ERROR RATES AND POWER LEVELS OF SELECTED MULTIVARIATE ANALYSIS OF VARIANCE PROCEDURES

Thesis Approved: William Thomas Coombs

Thesis Adviser

Katye M. Perry

Charles L Davis

Margaret M Scott

Thomas C. Collins

Dean of the Graduate College

ACKNOWLEDGMENTS

I wish to express my gratitude to the many people who have supported and assisted me throughout this project. First, I am indebted to Dr. William T. Coombs for (a) his guidance, his suggestions, and his own efforts to help me focus mine, (b) his unbounding energy in helping me tackle both philosophical and practical issues, and (c) his willingness to distinguish between persistence and nuisance. I was always inspired by his unswerving commitment to the best traditions in academic research, specifically those that preserve its integrity. To the other members of my committee, Dr. Dale Fuqua, Dr. Katye Perry, Dr. Charles Davis, and Dr. Margaret Scott, I wish to express thanks for reading countless pages of manuscript and offering encouragement, support, and counsel.

To my family and friends, especially husband Bob, daughter Betsy, mother Margaret Olson, and sister Susan Biegler, I apologize for the constant litany of "dissertation" talk you have endured over the last five years. I also want to thank you for your support and faith in me.

Finally, I would like to offer my appreciation to my friends and colleagues at Oral Roberts University and to the Administration for its continuous support. Specifically, I am thankful to Carole Todd and Dr. Thomas Luiskutty for assuming extra duties so I could complete my research. A special thanks to Russ Sowell for his help in keeping my link with OSU, my computer, always up and running and for his willingness to consult on computer issues at any time. I feel blessed to work with these people.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF SYMBOLS

α	nominal Type I error rate
τ	actual Type I error rate
β	Type II error rate
$1-\beta$	power of the test
H_0	the null hypothesis
$\underline{k}$	number of groups (also number of populations or number of samples)
Σ	population covariance matrix
$\underline{t}$	test statistic of independent samples $\underline{t}$-test
$\underline{F}$	test statistic of analysis of variance
$\underline{n}$	sample size
μ	population mean
σ^2	population variance
$\underline{t}_v$	test statistic of 2-sample Welch APDF test
$\underline{Q}$	test statistic of Wilcox test
$\underline{H}_M$	test statistic of Wilcox one-step M-estimator test
θ	variance ratio
$\underline{F}_v$	test statistic of $\underline{k}$-sample Welch APDF test
$\underline{J}$	test statistic of James test

$\underline{F}^*$	test statistic of Brown-Forsythe test
$\underline{H}$, $\underline{H}_m$	test statistics of Wilcox tests
$\underline{Z}$	test statistic of Wilcox $\underline{Z}$ test
$\underline{A}$	test statistic of Alexander and Govern test
$\underline{n}_r$	ratio of largest to smallest sample size
$\underline{N}$	total of sample sizes
$\underline{T}^2$	test statistic of Hotelling's test
$\mathbf{S}$	sample covariance matrix
$\mathbf{I}$	identity matrix
$\underline{p}$	number of dependent variables
$\underline{n}_{lg}$	largest sample size
$\underline{n}_{sm}$	smallest sample size
$\underline{T}_v^2$	test statistic of Nel and van der Merwe test
$\mathbf{H}$	hypothesis sum of squares and cross products matrix
$\mathbf{E}$	error sum of squares and cross products matrix
$\underline{R}$	Roy's largest root criterion
$\underline{U}$	Hotelling-Lawley trace criteria
$\underline{L}$	Wilk's likelihood ratio criterion
$\underline{V}$	Pillai-Bartlett trace criterion
$\underline{C}$	test statistic for Johansen test.
$\underline{R}^*$	Coombs-Algina $\underline{R}^*$ statistic.
$\underline{U}_1^*$	test statistic for Coombs-Algina $\underline{U}_1^*$ test.

$\underline{U_2}^*$	test statistic for Coombs-Algina $\underline{U_2}^*$ test.
$\underline{L}^*$	test statistic for Coombs-Algina $\underline{L}^*$ test.
$\underline{V}^*$	test statistic for Coombs-Algina $\underline{V}^*$ test.
$\underline{DT}$	distribution type.
$\underline{F}$	sample size ratio form.
$\underline{r}$	ratio between smallest sample size and number of dependent variables
$\underline{d}$	degree of heteroscedasticity.

Chapter 1
Introduction

When testing any statistical hypothesis the researcher has two concerns regarding the technical merits of the test employed: the test's significance level and its power. Both ideas are part of every introductory statistics course.

The significance level, traditionally denoted α, is the expected probability of a Type I error. That is, it is the probability a null hypothesis will be rejected when it is, in fact, true. Sometimes called the size of a test (especially in older literature), a Type I error is in a manner of speaking a false alarm--the chance something of significance will be discovered when it doesn't really exist. The significance level, then, tells the Type I error rate expected by the researcher. The actual Type I error rate, τ, may under certain conditions deviate from the expected or nominal Type I error rate, α. A smaller deviation is, of course, preferable to a larger one. A technically sound test is one in which τ is controlled so as to not deviate greatly from α. Tests in which τs are within allowable tolerances are termed robust. Studies that examine the Type I error rates of statistical tests, particularly under assumption violations, are called robustness studies. Such studies may be either analytic or empirical in nature.

The power of an omnibus hypothesis test is the test's ability to detect population differences that actually exist. Thus power, or sensitivity, is the probability that a false null hypothesis is rejected and, hence, a true difference is identified. High power is desirable. Power is traditionally denoted $1-\beta$, since it is the probability of the complement of a Type II error, the error of failing to reject a false null hypothesis, traditionally denoted β. Statistical studies that examine the power of hypothesis tests are called power analyses. Unlike robustness studies, power analyses are further complicated by the nature of the alternative to the null hypothesis. If the null hypothesis, H_0, is not true, something else must be true. Effect is a measure of the degree to which the alternative hypothesis differs from the null hypothesis. Both the size and form of the effect (or difference) are important. The power of a statistical test is a function of three factors: the significance level, the sample size, and the effect.

Robustness studies and power analyses are best done in concert, since the two ideas are inherently related. The concept of power is conditional upon a given significance level (Budescu & Appelbaum, 1981). So, power studies that do not first equate tests using significance level, are making comparisons on an unlevel playing field (Lee & Gurland, 1975).

Various conditions affect the ability of a statistical test to adequately control actual Type I error rate while maintaining sufficiently high power. The most evident are violations of the assumptions upon which the test is based. Common assumptions deal with relationships among samples (independent or dependent), distributions of underlying populations, and patterns of variability.

The Problem

One of the classic statistical tests is the comparison of the means of two populations. When random samples are independent and populations are normal in distribution and equal in variance, the solution has historically been the independent samples $\underline{t}$ test. According to Wang (1971), Behrens (1929) was the first to suggest a solution when the assumption of equal variances, or homoscedasticity, cannot be made, either because the variances are unknown or are known to differ. Fisher (1939) extended Behrens's solution, showing it to be the correct fiducial solution based on Fisher's theory of inference. The problem of testing the null hypothesis H_0: $\mu_1 = \mu_2$ under possibly heteroscedastic conditions became known as the Behrens-Fisher problem. Numerous parametric solutions to the Behrens-Fisher problem have been offered, including those by Welch (1947), Aspin (1948, 1949), Cochran and Cox (1950), Wald (1955), Yuen (1974), Lee and

Gurland (1975), and Wilcox (1992).

The analysis of variance (ANOVA) F test is a generalization of the independent samples t test to k samples. As in the two-sample case, independence of random samples, normality of distributions, and homoscedasticity ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$) are assumed. Proposed solutions in the absence of the homogeneity-of-variance (homoscedasticity) assumption are numerous. Among those that are parametric are the Welch approximate degrees of freedom (APDF) test (Welch, 1951), the James series tests (James, 1951), and the Brown-Forsythe test (Brown & Forsythe, 1974). Other parametric solutions to the k-sample Behrens-Fisher problem have been suggested by Marascuilo (1971), Rubin (1982), Wilcox (1988, 1989, 1993), and Alexander and Govern (1994).

Extending the independent samples t test to the multivariate case, Hotelling (1951) derived a test for the equality of two mean vectors, the case in which the null hypothesis is H_0: $\mu_1 = \mu_2$. The assumptions are multivariate analogs of those of the univariate test. Samples are assumed independent; populations are assumed multivariate normal; and population covariance matrices (also called dispersion matrices or variance-covariance matrices), Σ_1 and Σ_2, are assumed equal. Generalizations of univariate tests led to proposed solutions to the multivariate two-sample Behrens-Fisher problem by James (1954), Yao (1965), Johansen (1980), Nel and van der Merwe (1986), and Kim (1992).

Multivariate analysis of variance (MANOVA) is the extension of analysis of variance (ANOVA) to more than one dependent variable. It is also the extension of Hotelling's two-sample test to the k-sample case. Four classic MANOVA tests are based upon the works of Wilks (1932), Lawley (1938), Bartlett (1939), Roy (1945), Hotelling (1951), and Pillai (1955). Each tests for the equality of k population mean vectors (H_0: $\mu_1 = \mu_2 = \ \ldots \ = \mu_k$) under the assumption that k independent samples are randomly selected from k identically distributed multivariate normal populations. Specifically, all k covariance matrices are equal. Solutions suitable for use when covariance conditions are not known to be equal have been suggested by James (1954) who extended the work of James (1951) and Johansen (1980) who extended the work of Welch (1951). Coombs and Algina (in press) extended the Brown-Forsythe univariate test (Brown & Forsythe, 1974) to five tests that parallel the classic MANOVA procedures (two extensions of one test and one of each of the remaining three tests).

Purpose of the Study

The purpose of this study is to compare--under various experimental conditions--the Type I error rates and power levels of selected alternatives to the classic MANOVA procedures. The Pillai-Bartlett (Bartlett, 1939; Pillai, 1955) test, Johansen (1980) test, and four Coombs-Algina (in press) tests (one is omitted because of lack of a convenient F transformation) will be used to test H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ under varying distributions, numbers of groups, numbers of dependent variables, sample size ratio forms, ratios of smallest sample size to number of dependent variables, degrees of heteroscedasticity, and relationships between covariance matrices and sample sizes. Type I error rates will be computed and compared for the test statistics. To assess power two alternatives to the null hypothesis will be modeled in terms of effect size and form. Power will be computed and compared for those test statistics competitive in terms of Type I error control. Recommendations will be offered both on the basis of control of Type I error rate and power for competing tests.

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

Research Question 1. Does Type I error rate vary as a function of distribution type, number of groups, number of dependent variables, sample size ratio form, ratio of smallest sample size to number of dependent variables, degree of heteroscedasticity, or relationship between sample sizes and covariance matrices?

Research Question 2. Does power level vary as a function of distribution type, number of groups, number of dependent variables, sample size ratio form, ratio of smallest sample size to number of dependent variables, degree of heteroscedasticity, relationship between samples sizes and covariance matrices, or form of deviation from the null hypothesis?

Research Question 3. Under what conditions does each test maintain adequate control of Type I error rate and have suitable power?

Significance of the Study

The importance of the significance level of a hypothesis test has long been recognized. Tversky and Kahneman (1971) characterize statistical tests as protecting the scientific community by policing its members against overly hasty rejections of null hypotheses, in other words, against making Type I errors. More recently researchers have begun to include power as an important criterion upon which to base test selection and interpretation. Olson (1974) takes the view that a very high Type I error rate makes a test dangerous and that low power makes it useless. Stevens (1980) offers two reasons why power in inferential studies deserves a centerpiece roll:

> 1. High power (a priori) gives the researcher a reasonable chance of finding a difference if one exists. Surely if time, money, and resources are invested, one should demand a high chance of performing a successful inquiry, i. e., one that uncovers true differences.
>
> 2. A knowledge of power (post hoc) enhances the researcher's ability to correctly interpret nonsignificant results. Was a difference not discovered or

does it not exist? High power argues in favor of its nonexistence.

Researchers have been admonished periodically to pay more attention to power and less to statistical significance (Rossi, 1990). Further, attempts have been made to reduce power calculations to usable forms (Cohen, 1988, 1992; Koele, 1982). However, studies of the literature indicate that power of statistical tests has shown no notable increase in the last quarter century (Cohen, 1992; Freiman, Chalmers, Smith, & Kuebler, 1978; Moher, Dulberg, & Wells, 1994; Pulver, Bartho, & McGrath, 1988; Rossi, 1990). According to Cohen (1992) the absence of attention to power in the literature by both researchers and editors is inexplicable.

Multivariate tests are enjoying a dramatic increase in use in education and the behavioral sciences (Coombs, 1993). One reason may be their ability to provide greater power for rejecting a global null hypothesis than a collection of univariate tests, even when the multivariate tests are providing more stringent control over Type I error rate (Ramsey, 1982). Another reason is that educational research is inherently multivariate in nature, its outcomes seldom being measured against a single criterion variable (Stevens, 1972).

Hence, the goal of selecting a multivariate procedure that controls Type I error rate, while at the same time maintaining adequate power, is one of merit. Realization of this goal is seriously jeopardized when the assumptions upon which multivariate tests are based fail. Under inequality of population covariance matrices significance level is seriously affected for unequal sample sizes (Ito & Schull, 1964). Violations affect power adversely for both equal- and unequal-sized samples (Ito & Schull, 1964; Olson, 1974). Power is also negatively affected by numerous types of violations of multivariate normality, most notably kurtosis (Olson, 1974). The occurrence of non-normality in real-world data is common

(Cressie & Whitford, 1986; Micceri, 1989; Tiku, 1980). This sensitivity of the classic multivariate procedures to assumption violations, especially heteroscedastic conditions, in terms of both Type I error rate and power, (Korin, 1972; Olson, 1974; Pillai & Sudjana, 1975; Stevens, 1992), suggests that investigations of alternatives lacking such sensitivity are warranted. The James (1954) second-order test, Johansen (1980) test, and Coombs-Algina $\underline{R}^*$, $\underline{U}_1^*$, $\underline{U}_2^*$, $\underline{L}^*$, and $\underline{V}^*$ (in press) tests are just such alternatives. Researchers and practitioners alike will benefit from an illumination of the technical qualities of these alternatives to classic multivariate procedures.

This study examines only parametric tests. Although nonparametric alternatives are satisfactory answers to normality violations in the univariate case (Blair 1981, Zimmerman & Zumbo, 1993), they are sensitive to unequal variances just as parametric tests are (Zimmerman & Zumbo, 1993). Tomarken and Serlin (1986) found parametric approximations generally superior to nonparametric approaches in all but a few cases of nonnormality. Pratt (1964) and Tomarken and Serlin (1986) discourage their use as alternatives to the independent samples $\underline{t}$ test when variances differ. Blair (1981) favors nonparametric alternatives to the ANOVA $\underline{F}$ test based on their superior power when normality is violated, but argues against their use under heteroscedastic conditions because of the effect on Type I error properties. In the multivariate case, nonparametric tests are difficult to use owing to the complex and often enormous amount of computations required (Nath & Duran, 1983). Hence, the advantage of simple calculations present in the univariate case is lost in the multivariate case.

Chapter 2
Review of Literature

Independent Samples t Test

The independent samples t test is used to test the null hypothesis that two populations have the same mean under the assumptions that random samples are independent, both populations are normal in distribution, and both populations have equal variances. The operative test statistic

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

is distributed t with $n_1 + n_2 - 2$ degrees of freedom.

Many investigators have examined the independent samples t test both for Type I error rate robustness to assumption violations and for power. Early studies presented evidence showing the test to be nearly immune to assumption violations other than independence, which was explored by Daniels (1938). These studies concluded that neither departures from normality (Box & Andersen, 1955; Bartlett, 1935; Cochran, 1947; David & Johnson, 1951; Gayen, 1950; Neave & Granger, 1968; Pearson, 1931) nor departures from homoscedastic conditions (Box, 1954; Horsnell, 1953; Welch, 1938) caused the actual Type I error rate τ to differ greatly from the nominal rate α, unless the departures were so severe as to be readily apparent upon a mere inspection of the data (Lindquist, 1953).

Boneau (1960) and Havlicek and Peterson (1974) extended these studies and found the effect of unequal variances upon Type I error rate to be related to sample size. Under direct pairing, when the larger sample comes from the population with the larger variance (the positive condition), the t test is conservative; that is, $\tau < \alpha$. On the other hand the test is liberal, $\tau > \alpha$, in the negative condition, when the larger sample comes from the population with the smaller variance. Pratt (1964) examined the behavior of τ mathematically for the case in which the sample size ratio is 3:2. The variance ratio θ was allowed to vary across its entire range $(0, \infty)$ yielding Type I error rates that varied from .016 to .109 for $\alpha = .05$.

Ramsey (1980) applied Hsu's (1938a) equations to demonstrate that even when sample sizes are the same, the t test is not always robust to violations of homoscedasticity. Using Bradley's liberal criterion for robustness (Bradley, 1978) and Cochran's limits for robustness (Cochran, 1954), Ramsey offered equal-sample-size guidelines for t test robustness at various levels of significance. Ramsey concluded by suggesting the use of alternative statistics when these guidelines cannot be followed, especially when sample sizes are not equal.

When the assumptions are satisfied, the t test is uniformly the most powerful among the unbiased size α tests for the significance of the difference between two means (Best & Rayner, 1987; Blair, 1981). While t does retain its power under some assumption violations (David & Johnson, 1951), often the power of the test is affected by departures from normality. Power tends to increase (though actual Type I error rate does not) when both populations are skewed in the same direction. When the populations are skewed in opposite directions, the power function is markedly distorted (Young & Veldman, 1963). If both populations are symmetric and samples equal, nonnormality has little effect on either Type I error control or power (Tan, 1982). The effect of kurtosis on power is greater than that of skewness, but equal sample sizes tend to diminish the effects of the lack of normality on power (Pearson, 1929; Tan, 1982). Violations of the assumption of homogeneity of variance have little influence on the power of the t test (Young & Veldman, 1963).

Neave and Granger (1968) compared the $\underline{t}$ test and seven nonparametric alternatives for power under the following conditions: (a) samples were selected from both normal and nonsymmetric, bimodal distributions, (b) the variance ratio $\theta = \sqrt{2}$ or 1, and (c) $\underline{n}_1 = \underline{n}_2 = 20$ or $\underline{n}_1 = 20$ and $\underline{n}_2 = 40$. Power was calculated for mean differences of $\frac{1}{2}$ and 1. The authors found power levels to improve with increasing sample size, being higher for sample sizes of 20 and 40 than for equal sample sizes of 20. In the Neave and Granger study power levels tended to be slightly lower under heteroscedastic conditions, but on balance estimated power levels agreed quite well with theoretical levels. Departures from normality in the form of nonsymmetric bimodality did not appreciably affect power.

Donaldson (1968) examined the effect of heteroscedasticity on the power of the $\underline{t}$ test under the following conditions: (a) samples were selected from normal, exponential, and lognormal distributions, (b) the variance ratio $\theta = 1, 1.56, 2.25,$ or 12.25, and (c) $\underline{n}_1 = \underline{n}_2 = 16$. Effect size expressed as

$$\sqrt{\frac{16\sum_{i=1}^{2}(\mu_i-\overline{\mu})^2}{2\,\overline{\sigma}^2}}\,,$$

where $\overline{\mu}$ and $\overline{\sigma}^2$ are the average mean and variance, ranged from .44 to 1.94. Donaldson found that when effect size was very small, the test performed using samples selected from normal distributions displayed slight power advantages over the test performed using samples from either the exponential or lognormal distributions. The situation quickly reversed as effect size increased. For both the exponential and lognormal distributions, tests performed under heteroscedastic conditions showed smaller differences in power for small effects and larger differences in power for large effects when compared to the same tests performed under homoscedasticity. Under normality only slight differences, attributable to sampling error, occurred between the unequal- and equal-variance cases. Donaldson concluded that under normality, there is a close correspondence between actual and nominal power levels.

Alternatives to the Independent Samples t Test

Wang (1971) reported that the first exact solution to the problem of testing for the difference between the means of two populations with unknown variances was supplied by Behrens (1929) and extended by Fisher (1939) as the correct fiducial solution. Fisher (1935), Sukhatme (1938), and Fisher and Healy (1956) calculated tables for the distribution of the Behrens-Fisher statistic.

Welch (1938) identified two statistics used to test for equality of means in the absence of equal variances, which he called u and v. When sample sizes are equal, u and v are equal. Several studies (Fenstad, 1983; Gronow, 1951; Welch, 1938) established the superiority of the Welch v. The statistic, which is often denoted t_v, is

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

Several tests have been proposed using the Welch v statistic or t_v. Since the statistic does not yield an exact test (Welch, 1938), all solutions fall into one of two categories: (a) approximate degrees of freedom (APDF) solutions and (b) series solutions. The APDF tests are derived by approximating degrees of freedom which define the sampling distribution (in this case a t distribution with v degrees of freedom). Series solutions are derived by utilizing a series expansion to determine the critical value for the rejection region.

Welch (1947) showed that t_v follows a t distribution with degrees of freedom

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

The value of f is not necessarily an integer. The test using this estimator is known as the Welch APDF test.

Welch (1947) developed a series expression for the critical value of t_v as a function of the significance level, sample sizes, and sample variances. The critical values for the zero-, first-, and second-order series solutions appear in Table 1. Aspin (1948) calculated third- and fourth-order solutions.

Yuen (1974) suggested a two-sample test based on trimmed means and Winsorized variances whose critical value is a percentile of the Student's t distribution. The test was developed from the work of Yuen and Dixon (1973) and is commonly referred to as Yuen's trimmed means test or Yuen's trimmed t test.

Wilcox (1989) proposed a modification of the zero-order Welch series solution using the statistic

$$Q = \frac{\tilde{x}_1 - \tilde{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} ,$$

where

$$\tilde{x}_i = \frac{2x_{in_i}}{n_i(n_i + 1)} + \frac{n_i - 1}{n_i(n_i + 1)} \bar{x}_i .$$

Wilcox's statistic is approximately normal in distribution. The estimators $\tilde{x}_1$ and $\tilde{x}_2$ for μ_1 and μ_2 are biased for the purpose of reducing the difference between the actual and nominal Type I error rates.

Table 1

Critical Values for Welch's (1947) Zero-, First-, and Second-Order Series Solutions

Order	Critical Value
Zero	z
One	$z\left(\frac{1+z^2}{4}\,\frac{\sum_{i=1}^{2}\frac{\left(\frac{s_i^2}{n_i}\right)^2}{n_i-1}}{\left(\sum_{i=1}^{2}\frac{s_i^2}{n_i}\right)^2}\right)^{\frac{1}{2}}$
Two	$z\left(-\frac{1+z^2}{4}\,\frac{\sum_{i=1}^{2}\left(\frac{s_i^2}{n_i-1}\right)^2}{\left(\sum_{i=1}^{2}\frac{s_i^2}{n_i}\right)^2}\right.$ $+\frac{3+5z^2+z^4}{3}\,\frac{\sum_{i=1}^{2}\left(\frac{\left(\frac{s_i^2}{n_i}\right)^3}{n_i-1}\right)^2}{\left(\sum_{i=1}^{2}\frac{s_i^2}{n_i}\right)^3}$ $\left.-\frac{15+32z^2+9z^4}{32}\,\frac{\sum_{i=1}^{2}\left(\frac{\left(\frac{s_i^2}{n_i}\right)^2}{n_i-1}\right)^2}{\left(\sum_{i=1}^{2}\frac{s_i^2}{n_i}\right)^4}\right)^{\frac{1}{2}}$

Note: z is a percentile of the standard normal distribution.

Wilcox (1992) suggested a statistic based on one-step M-estimators of location:

$$H_M = \frac{\bar{x}_{m1} - \bar{x}_{m2}}{\sqrt{s^2_{m1} + s^2_{m2}}} \quad .$$

In H_M, $\bar{x}_{mi}$ is the one-step M-estimator in the ith group and s_{mi} is the estimated standard error of the ith estimator. The test using this statistic employs bootstrap methods to calculate the critical value. This Wilcox test and the Yuen trimmed means test were developed as ways to deal with heavy tails and outliers (Wilcox, 1992) which have substantial effects on power (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Yuen, 1974).

Various other statistics have been proposed to replace the independent samples t test when variances differ. One such test is based on modified maximum likelihood estimators of location and scale parameters of symmetric distributions (Tiku and Singh, 1981; Tiku, 1980, 1982). Wald (1955), Cochran and Cox (1950), and Lee and Gurland (1975) have proposed others.

The literature suggests the following regarding the independent samples t test; Welch APDF test; the Welch zero-, first-, and second-order series tests; the Aspin third- and fourth-order tests; the Yuen trimmed means test; the Wilcox Q test; and the Wilcox H_M test in terms of the control of Type I error rate under heteroscedasticity: (a) the Welch APDF and James second-order tests are superior to the James first-order test which is superior to the independent samples t test, (b) both the Wilcox and the Welch second-order series tests are superior to the Welch APDF test, (c) the Aspin test is only slightly superior to the Welch APDF test, (d) the Welch APDF test is adequate under normality, and (e) the James second-order and Wilcox H_M tests best control Type I error rate under the widest range of conditions. In terms of power the literature suggests: (a) the Welch APDF test is more powerful in the positive condition and (b) the Welch-Aspin series and Welch APDF tests are comparable in power, (c) little power is lost using the Welch APDF test in place of the independent samples $\underline{t}$ test, even when assumptions are satisfied, (d) the Welch second-order series test is slightly more powerful than the Wilcox $\underline{Q}$ test, and (e) the Wilcox $\underline{H}_M$ test is more powerful than either the Welch APDF test or the Yuen trimmed means test.

Gronow (1951) studied the Welch APDF test in terms of power when variances are unequal. Heteroscedasticity somewhat reduced power for $\underline{n}_1 = \underline{n}_2 = 10$. For unequal sample sizes, power was found to be higher in the positive condition. Actual Type I error rate, however, sometimes varied greatly from the nominal rate.

Scheffé (1970) considered six solutions to the Behrens-Fisher problem and provided mathematical approximations of power for those he preferred, among them the Welch APDF test and the Welch-Aspin series solutions. Scheffé found the Welch-Aspin series tests to be more powerful than the Behrens-Fisher test and superior in Type I error control. He also found the power of the Welch-Aspin series tests to be well approximated by the power of the Welch APDF test. His comparisons lead to the conclusion that Welch's APDF test is a practical solution to the Behrens-Fisher problem, despite a slight disadvantage in Type I error rate, because it requires only the "ubiquitous" $\underline{t}$ tables.

Wang (1971) compared the Behrens-Fisher test, Welch APDF test, and Aspin series tests. She found that differences between actual and nominal Type I error rates were small for the Welch APDF test under heteroscedastic conditions, the largest deviation being .0035 with samples of sizes 5 and 21 at $\alpha = .01$. The Aspin series tests had slightly smaller deviations. The Behrens-Fisher test was found to be quite conservative, while the Welch APDF test was found to be slightly liberal. All things considered, especially the lack of availability of Aspin series values and the tedious nature of the computations for the series tests, Wang (as Scheffé) recommended the Welch APDF test.

Yuen (1974) compared the Yuen trimmed means and Welch APDF tests under normality and long-tailedness in a Monte Carlo experiment using both equal and unequal sample sizes selected from populations with different variances. She found both tests to be conservative under long-tailedness, the Welch APDF test more so. The trimmed means test was found to be generally superior in terms of power, with the level of superiority depending on degree of long-tailedness, sample sizes, and level of mean trimming. Under normality the Yuen test had lower power levels than the independent samples $\underline{t}$ test, although the power loss was small when the amount of trimming was small.

Hampel, Ronchetti, Rousseeuw, and Stahel (1986) reported that both the independent samples $\underline{t}$ test and the Welch APDF test are considered in most cases to be robust to assumption violations in terms of Type I error rate, but not in terms of power.

Acknowledging that there is no uniformly most powerful unbiased size α test for the Behrens-Fisher problem for all sample sizes, Best and Rayner (1987) recommended the routine use of the Welch APDF test, regardless of whether assumptions are satisfied or not. They based this recommendation on the results of simulations under the following conditions: (a) samples were taken from normal distributions, (b) the variance ratio $\theta = \frac{1}{2}, \frac{1}{4}$, 1, 2, or 4, (c) ($\underline{n}_1$, $\underline{n}_2$) = (4, 8), (5, 15), (10, 10), (15, 45), (30, 30), (25, 75). Power comparisons were made for effect sizes in which the population mean differences were 1, 2, 3, or 4 standard errors. Best and Rayner found the Welch APDF test to perform well in terms of power in all conditions. The effect of heteroscedasticity diminished with increased sample sizes. Even when the variance ratio $\theta = 1$, the loss in power of the Welch APDF test compared to the independent samples $\underline{t}$ test was of no practical importance for degrees of freedom of at least 5.

Wilcox (1989) found James's second-order series test, which reduces to Welch's second-order series test when $\underline{k} = 2$, to control Type I error rate almost as well as the Wilcox test and to have slightly more power. For $\underline{k} = 2$ the advantage of the Wilcox test over the Welch APDF test was very slight when both populations were normal.

Wilcox (1990) studied five tests under departures from both normality and homoscedasticity, among them the Wilcox $\underline{Q}$ and Welch APDF tests. Monte Carlo conditions included: (a) sampling from two populations, the first of which was usually normal and the second of which came from a distribution with one of five levels of skewness or one of five levels of kurtosis, (b) variance ratio $\theta = 1, 2,$ or 4, (c) $\underline{n}_1 = 12, 20, 30, 40, 60,$ or 80 and $\underline{n}_2 = 12$ or 20. Power was studied by adding a constant to every observation in the second group. Wilcox found the Welch APDF test to be robust in terms of Type I error rate to both violations when $\underline{n}_1 = \underline{n}_2$. When sample sizes differed, it was sometimes too liberal. The Wilcox test was affected least by departures from normality and it, too, was found to be liberal in some conditions. The power of both tests was affected by departures from normality. When samples sizes were equal, Wilcox found the Welch APDF test to be superior to the Wilcox $\underline{Q}$ test in terms of both control of Type I error rate and power. When sample sizes differed, the Wilcox test was superior to the Welch APDF test in both control of Type I error rate and power. Under assumption violations, Wilcox recommended using the Welch APDF test when sample sizes were the same and the Wilcox $\underline{Q}$ test when they differed.

Wilcox (1992) compared the Wilcox $\underline{H}_M$ test with the Yuen trimmed mean and Welch APDF tests. His simulation study found the $\underline{H}_M$ test to have stable Type I error control ($.031 \leq \tau \leq .055$) for tests of noncontaminated identically shaped distributions in which the variance ratio was 4:1 or less. Good control was also exhibited when population shapes differed. Five distribution types were included in the study: normal, $\underline{t}$ with 5 degrees of freedom, exponential, moderate skewness, and extreme nonnormality. In terms of power $\underline{H}_M$ was found to be substantially more powerful than the trimmed means test which in turn outperformed the Welch APDF test as contamination increased.

Zimmerman and Zumbo (1993) examined the independent samples $\underline{t}$ test, Welch APDF test, and two nonparametric tests under simulated conditions in which: (a) random samples were selected from normal populations, (b) the variance ratio $\theta = 1$ or 16, and (c) sample sizes were 6, 12, or 18. For equal-sized samples and variance ratios $\theta = 1$ the powers of the $\underline{t}$ test and the Welch APDF test were similar. The Welch APDF test, however, exhibited power superior to that of the $\underline{t}$ test when $\underline{n}_1 = \underline{n}_2 = 18$ and the variance ratio $\theta = 16$.

In summary, the independent samples $\underline{t}$ test is generally acceptable in controlling Type I error rate and is the uniformly most powerful unbiased size α test when all assumptions are met. Under heteroscedastic conditions, it remains acceptable for sufficiently large equal-sized random samples. However, when variances differ, superior alternatives exist that do not require equality of variances. The Welch APDF test seems to be the most practical solution, given its control of Type I error rate, acceptable power, and reliance on the easily accessible $\underline{t}$ distribution. Somewhat superior series solutions are available in the James second-order and the Wilcox $\underline{H}_M$ tests, but they are either computationally intense or unavailable in standard computer packages.

Analysis of Variance

The ANOVA F statistic tests for differences among the means of *k* independent samples randomly selected from *k* normally distributed populations with equal variances. The statistic

$$F = \frac{\dfrac{\sum_{i=1}^{k} n_{i.}(\bar{x}_i - \bar{x}_{..})^2}{k-1}}{\dfrac{\sum_{i=1}^{k} (n_i - 1)s_i^2}{N-k}}$$

is distributed *F* with ($k - 1$) and ($N - k$) degrees of freedom where $N = \sum_{i=1}^{k} n_i$. In the two-sample case the *F* statistic reduces to the square of the *t* statistic.

Numerous studies have concluded that violations of the assumption of equal variances affect the Type I error rate in the one-way analysis of variance. Early works by Box (1954) and Horsnell (1953) demonstrated mathematically that the ANOVA *F* test is robust to heteroscedasticity as long as sample sizes are equal. But these early studies considered primarily conditions in which θ, the ratio of the largest to smallest population variance was small, equal to $\sqrt{3}$. Box found one case that was a distortion of the idea that equal sample sizes eliminate the effects of unequal variances: an actual Type I error rate $\tau = .12$ with a nominal rate $\alpha = .05$, $\theta = \sqrt{7}$, and equal sample sizes of 3. Later studies extended Box's work in the direction of this "distortion" by considering conditions in which $\theta \geq \sqrt{3}$. These studies showed that for sufficiently large θ, the ANOVA *F* test is not robust to equal variance violations, even for equal sample sizes (Brown & Forsythe, 1974; Clinch & Keselman, 1982; Harwell, Rubinstein, Hayes, & Olds, 1992; Rogan & Keselman, 1977). Further, Fenstad (1983) and Wilcox (1987) have demonstrated that values of θ as large as 4 are not unusual in the literature.

When sample sizes are not equal, the ANOVA *F* test is conservative in the positive condition and liberal in the negative condition (Clinch & Keselman, 1982; Tomarken & Serlin, 1986). Because the ANOVA *F* test is more sensitive to unequal variances than previously thought and no test exists with adequate power to identify cases of heterogeneity, some investigators have suggested that researchers abandon the analysis of variance *F* test (Wilcox, 1987; Wilcox, Charlin, & Thompson, 1986).

The power of the ANOVA *F* test, both under assumptions and in the face of violations, has been studied extensively. McFatter and Gollab (1986) chronicled a short list of investigators. According to Glass, Peckham, and Sanders (1972) Horsnell (1953) produced the first published investigation of the *F* test under heteroscedastic conditions. Scheffé (1959) noted that as late as 1959 most of what was known regarding the effect of unequal variances on the power of the *F* test could be traced to the Horsnell study. Donaldson's (1968) empirical study, unlike Horsnell's analytic one, was restricted to equal sample sizes. Both concluded, however, that under normality a close correspondence exists between empirical or actual power and theoretical power calculated using a mean variance for the common variance required in theory to compute power.

Budescu (1982) conducted an empirical test of the power of the ANOVA *F* test under the following conditions: (a) samples were selected from normal distributions, (b) $\underline{k} = 4$, (c) samples were both equal and unequal in size, and (d) variances were proportional to means. Powers were calculated using three noncentrality parameters which described the amount of difference among the means. In addition, two forms of noncentrality, concentrated $(\mu_1 = \mu_2 = \mu_3 \leq \mu_4)$ and diffuse $(\mu_1 - 3\underline{a} \leq \mu_2 - 2\underline{a} \leq \mu_3 - \underline{a} < \mu_4)$, described mean configuration. Budescu concluded that the power of the ANOVA *F* test under normality with variances proportional to means can be well approximated by the normal power function using an estimated noncentrality parameter.

A number of investigators have discussed the power of the ANOVA F test under normality violations (Boneau, 1960; Donaldson, 1968; Games & Lucas, 1966; Srivastava, 1959; Tan, 1982; Tiku, 1971). Srivastava (1959) provided tables of powers for various values of skewness and kurtosis for various effect sizes and mean arrangements and demonstrated that increasing sample size decreases the effect of kurtosis on the power curve. Boneau (1960) showed that platykurtosis reduces power and leptokurtosis causes it to increase.

Games and Lucas (1966) conducted a Monte Carlo power analysis using three populations, samples of size 3 or 6, and nine distributions, eight of which were not normal. They found that for non-normal populations theoretical normal-theory power calculations were remarkably well approximated by the empirical power values in their simulations. However, while moderate departures from normality had little practical effect on power, extreme skewness and moderate leptokurtosis did produce great effects in power values.

Tiku (1971) also found moderate departures from normality to have little effect on the power of the F test. He noted that as sample size increases, kurtosis has a greater effect on power than skewness.

Tan (1982) summarized what is known about the effect of non-normality on the power of the F test. For moderate departures and equal sample sizes F is quite robust with respect to power. Severe departures (exponential and lognormal) exert considerable effects which become more pronounced as sample size differences increase. Kurtosis has a more dominant role than skewness in determining power.

Alternatives to the Analysis of Variance

Several alternatives to the ANOVA *F* statistic have been derived to test the null hypothesis of the equality of *k* means when the assumption of homoscedasticity does not hold; that is, when $\sigma_i \neq \sigma_j$ for at least one pair of *i* and *j*.

Welch (1951) extended the Welch (1947) APDF solution which yielded the statistic

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

where

$$w_i = \left(\frac{s_i^2}{n_i}\right)^{-1}, \qquad w = \sum_{i=1}^{k} w_i, \qquad \bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}\,, \qquad \text{and} \qquad \bar{x} = \sum_{i=1}^{k}\frac{w_i\bar{x}_i}{w} \quad .$$

The Welch APDF statistic F_V is approximately distributed as *F* with

$$k-1 \qquad \text{and} \qquad \left(\frac{3}{k^2-1}\sum_{i=1}^{k}\left(1-\frac{w_i}{w}\right)^2\right)^{-1}$$

degrees of freedom.

Marascuilo (1971) suggested a variation of the Welch APDF test that yields slightly larger values of the test statistic.

James (1951) proposed generalizations of the Welch (1947) series solutions. James's statistic is

$$J = \sum_{i=1}^{k} w_i(\bar{x}_i - \bar{x})^2$$

where

$$w_i = \left(\frac{s_i^2}{n_i}\right)^{-1}, \qquad \bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}\,, \qquad \bar{x} = \sum_{i=1}^{k} \frac{w_i \bar{x}_i}{w}\,, \quad \text{and} \qquad w = \sum_{i=1}^{k} w_i\,.$$

The statistic $\underline{J}$ is asymptotically χ^2 in distribution with $\underline{k} - 1$ degrees of freedom. However, the chi-square approximation is not satisfactory if sample sizes are small or even moderately large. James offered two methods for adjusting critical values in such cases, yielding what are known as the James first- and second-order solutions. In the first-order test, if all population means are equal,

$$P\,[\![J \geq \chi^2_{k\text{-}1;\alpha}\left(1 + \frac{3\chi_{k\text{-}1;\alpha} + k + 1}{2(k^2 - 1)} \sum_{i=1}^{k} \frac{1}{f_1}\left(1 - \frac{w_i}{w}\right)^2\right)]\!] = \alpha\,.$$

The James second-order method also yields approximate critical values, but is computationally challenging as observed by James.

Brown-Forsythe (1974) offered for consideration the test statistic

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

$\underline{F}^*$ is approximately distributed as $\underline{F}$ with $\underline{k} - 1$ and $\underline{f}$ degrees of freedom, where

$$f = \frac{\left(\sum_{i=1}^{k}\left(1 - \frac{n_i}{N}\right) s_i^2\right)^2}{\sum_{i=1}^{k} \frac{\left(\left(1 - \frac{n_i}{N}\right) s_i^2\right)^2}{n_i - 1}}\,.$$

Both the Welch (1951) APDF test and the Brown-Forsythe test reduce to the Welch (1947) APDF test when $\underline{k} = 2$.

Rubin (1982) modified the Brown-Forsythe test by utilizing the same test statistic, but substituting $\underline{g}_1$ for the numerator degrees of freedom, where

$$g_1 = [\sum_{i=1}^{k} \left(1 - \frac{n_i}{N}\right) s_i^2]^2 [[\sum_{i=1}^{k} \frac{n_i}{N} s_i^2]^2 + \sum_{i=1}^{k} \left(1 - 2\frac{n_i}{N}\right) s_i^2]^{-1}.$$

The denominator degrees of freedom $\underline{g}_2$ equals that of the Brown-Forsythe statistic which is a percentile of the $\underline{F}$ distribution with $\underline{g}_1$ and $\underline{g}_2$ degrees of freedom.

Wilcox (1988) proposed the statistic

$$H = \frac{\sum_{i=1}^{k} (w_i - \overline{w})^2}{\max_{i=1}^{k} \left(\frac{s_i^2}{n_i}\right)}$$

followed (1989) with a modified statistic

$$H_m = \sum_{i=1}^{k} w_i (\tilde{x}_i - \tilde{x})^2$$

where

$$w_i = \left(\frac{s_i^2}{n_i}\right)^{-1}, \quad w = \sum_{i=1}^{k} w_i, \quad \tilde{x}_i = \frac{2x_{in_i}}{n_i(n_i + 1)} + \frac{n_i - 1}{n_i(n_i + 1)} \overline{x}_i, \text{ and } \tilde{x} = \sum_{i=1}^{k} \frac{w_i \tilde{x}_i}{w}.$$

Both $\underline{H}$ and $\underline{H}_m$ are approximately distributed as χ^2 with $\underline{k} - 1$ degrees of freedom.

Wilcox (1993) proposed yet another statistic by generalizing his $\underline{H}_M$ test to the case of k groups:

$$Z = \frac{1}{N} \sum_{i=1}^{k} n_i(\bar{x}_{mi} - \bar{x}_m)$$

where

$$\bar{x}_m = \frac{1}{k} \sum_{i=1}^{k} \bar{x}_{mi} \qquad \text{and} \qquad N = \sum_{i=1}^{k} n_i \,.$$

The critical value for Z is calculated using bootstrap methods as with the two-sample case.

Alexander and Govern (1994) derived the test statistic

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

where

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

$$a = n_i - 1.5 \qquad b = 48a^2 \qquad c = [a \ln\left(1 + \frac{t_i^2}{n_i - 1}\right)]^{\frac{1}{2}}$$

$$t_i = \frac{\bar{x}_i - x^+}{s_i} \qquad x^+ = \sum_{i=1}^{k} w_i \bar{x}_i \qquad w_i = \frac{\frac{1}{s_i^2}}{\sum_{i=1}^{k} \frac{1}{s_i^2}}$$

and

$$s_i = \left(\frac{\sum_{j=1}^{n_i} (x_j - \bar{x}_i)^2}{n_i(n_i - 1)} \right)^{\frac{1}{2}} \,.$$

The Alexander-Govern statistic A is distributed approximately as χ^2 with k − 1 degrees of freedom.

The literature includes a number of conclusions regarding the control of Type I error rate and power for the ANOVA $\underline{F}$, Welch APDF $\underline{F}_v$, James first-order $\underline{J}$, James second-order $\underline{J}$, Brown-Forsythe $\underline{F}^*$, Wilcox $\underline{H}$, Wilcox $\underline{H}_m$, Wilcox $\underline{Z}$, and Alexander-Govern $\underline{A}$ tests under violations of homoscedasticity. Regarding Type I error control, conclusions suggest: (a) each of the alternatives is superior to the ANOVA $\underline{F}$, (b) both the Welch APDF and Brown-Forsythe tests outperform the James first-order test, (c) the Welch APDF and Brown-Forsythe tests are generally competitive with the Welch test enjoying a slight edge under normality, (d) the Rubin test is competitive with the Welch APDF test and outperforms the Brown-Forsythe test, and (e) the James second-order, Wilcox $\underline{Z}$, and Alexander-Govern $\underline{A}$ tests outperform all contenders.

Generally, it can be concluded regarding power that (a) the ANOVA $\underline{F}$ is the most powerful when variances are equal, (b) heteroscedasticity has a negligible effect on the ANOVA $\underline{F}$ test, (c) little power is lost when using the Welch APDF, Brown-Forsythe, James second-order, or Wilcox $\underline{H}_m$ tests, (d) the Wilcox $\underline{H}$ test has inferior power levels when compared to any of the other tests, (e) the Wilcox $\underline{H}_m$ test has slightly less power than the James second-order test, and (f) the Alexander-Govern $\underline{A}$ test and the James second-order tests have comparable power levels. In nearly all studies power comparisons were made only for tests showing adequate control of Type I error rates. The concept of power is conditional upon a given probability of Type I error. It must be shown that two tests operate at the same α level before power comparisons can be meaningful (Budescu & Appelbaum, 1981).

Brown and Forsythe (1974) examined the ANOVA $\underline{F}$, Brown-Forsythe $\underline{F}^*$, Welch APDF $\underline{F}_v$, and James first-order test via Monte Carlo experiments under the following conditions: (a) samples were selected from normal distributions, (b)

$\underline{k} = 4, 6,$ or 10, (c) the ratio of the largest to smallest population variance $\theta = 1$ or 3, (d) the ratio of the largest to smallest sample size $\underline{n}_r = 1, 1.9,$ or 3, and (e) 16 $\leq \underline{N} \leq$ 200 where $\underline{N} = \sum_{i=1}^{k} \underline{n}_i$ is total sample size. For power analyses three noncentrality structures were used with each of three variance configurations at both the .01 and .05 significance levels. Brown and Forsythe concluded that for both the Welch APDF and Brown-Forsythe tests, actual Type I error rate τ closely approximated the nominal rate α, while τ fluctuated greatly for the ANOVA $\underline{F}$. For small samples the critical value in the Welch APDF test was a better approximation to the true value than was the James first-order approximation. For that reason, the James test was omitted from power comparisons. The Welch APDF test was more powerful than the Brown-Forsythe test when extreme means were paired with small variances. When extreme means were paired with large variances, the Brown-Forsythe test was more powerful. Both the Welch $\underline{F}_v$ and the Brown-Forsythe $\underline{F}^*$ tests were only slightly less powerful than the ANOVA $\underline{F}$, even when homoscedasticity held.

Kohr and Games (1974) studied the ANOVA $\underline{F}$ test, Box test, and Welch APDF test under the following conditions: (a) samples were selected from normal distributions, (b) $\underline{k} = 4$, (c) largest to smallest population variance ratio $\theta = 1, \sqrt{2}, \sqrt{7}, \sqrt{10}$, or $\sqrt{13}$, (d) largest to smallest sample size ratio $\underline{n}_r = 1, 1.5,$ or 2.8, (e) $\underline{N} =$ 32 or 36. Power analyses were made for 47 of 81 possible combinations of 9 variance conditions, 3 noncentrality structures, and 3 sample size conditions using $\alpha = .05$. Kohr and Games concluded that the Welch APDF test exhibited better control of Type I error rates under heteroscedasticity than either the ANOVA $\underline{F}$ test or the Box test, although the superiority was more pronounced when compared to the ANOVA $\underline{F}$ test. Of the three tests the Box procedure was never the most powerful. For equal or unequal sample sizes and homogeneous variances the ANOVA F test was more powerful than the Welch APDF test. Under heteroscedasticity and equal sample sizes the ANOVA F test was more powerful when extreme means were paired with larger variances. However, under these conditions the ANOVA F test did not adequately control Type I error rates, so the comparison lacked validity. The Welch APDF test was most powerful when sample sizes and variances varied.

Levy (1978b) compared empirically the ANOVA F test, the Welch APDF F_V test, and the Marascuilo variation of the Welch APDF test, examining both Type I error control and power. Conditions studied in the robustness test included: (a) samples were selected from uniform, double exponential, χ^2 (5 degrees of freedom), and exponential distributions, (b) k = 3 or 6, (c) largest to smallest sample size ratio n_r = 1, 3, or 5 for k = 3 and n_r = 1, 3, 1.7, and 2.5 for k = 6, and (d) largest to smallest population variance ratio θ = 1 or 50. Power was studied only under homogeneity of variances. The ANOVA F and Welch APDF tests were comparable in controlling Type I error rate under equal variances. The Marascuilo test was liberal except when all sample sizes were at least 15. Under heteroscedasticity and equal-sample-size conditions, F was liberal, F_V was adequate, and the Marascuilo statistic was somewhat liberal. In the positive condition the ANOVA F was conservative, while actual and nominal Type I error rates were approximately equal ($\tau \approx \alpha$) for the other two tests. In the negative condition F was liberal, while the other two tests were again satisfactory. In terms of Type I error control, the Welch APDF test was the best of the three studied under heteroscedasticity and was comparable to the ANOVA F under homoscedasticity. The Marascuilo test is inherently more powerful than the Welch APDF test, since it always yields slightly larger values of the test statistic. Levy's simulation showed the ANOVA F to be slightly more powerful than the

Welch APDF test. It was also generally more powerful than the Marascuilo test except for two sample-size instances ($\underline{k} = 3$ and $\underline{n}_i = 5$; $\underline{k} = 6$ and $\underline{n}_i \leq 15$). All tests were robust to nonnormality with $\underline{F}$ showing the highest degree of robustness followed by the Welch APDF and Marascuilo tests in that order. In a separate study Levy (1978a) showed that the non-normal distribution of the Welch APDF statistic can be approximated by an approximate noncentral $\underline{F}$ distribution and demonstrated the closeness of the approximation using Monte Carlo simulations with conditions similar to those in the Levy (1978b) study.

Dijkstra and Werter (1981) considered the Welch APDF test, Brown-Forsythe test, and James second-order test empirically under the following conditions: (a) samples were drawn from normal populations, (b) $\underline{k} = 3$, 4, or 6, (c) largest to smallest population variance ratio $\theta = 1$ or 3, (d) largest to smallest sample size ratio $\underline{n}_r = 1$, 2, 2.5, or 3.5, and (e) $12 \leq \underline{N} \leq 90$. Power analyses were conducted using four noncentrality structures, two variance configurations, and two sample-size conditions at significance levels of .01, .05, and .10. The James second-order test gave better protection against unequal variance effects on Type I error rates than either the Brown-Forsythe or the Welch APDF tests, which were comparable. None of the three tests was uniformly more powerful than the other two. The Brown-Forsythe test was more powerful when extreme means coincided with larger variances, while the Welch APDF test and James second-order test had power advantages when extreme means coincided with small variances.

Clinch and Keselman (1982) compared the ANOVA $\underline{F}$ test, the Welch APDF test, and the Brown-Forsythe $\underline{F}^*$ test under the following conditions: (a) samples were taken from distributions that were normal, χ^2 with 2 degrees of freedom, or $\underline{t}$ with 5 degrees of freedom, (b) $\underline{k} = 4$, (c) largest to smallest population variance ratio $\theta = 1$, 1.32, 1.82, 3.04, or 4.22, (d) largest to smallest sample size $\underline{n}_r = 1$ or 3, and (e) $\underline{N} = 48$ or 144. Two alternative distributions of means were examined for power comparisons using the .05 significance level and three sample size-variance pairings (equal sample sizes, direct pairing, and inverse pairing). When sampling was from symmetric distributions, the Welch APDF and Brown-Forsythe tests exhibited adequate Type I error control, both tests withstanding the combined effects of unequal group sizes and variance heterogeneity. The ANOVA $\underline{F}$ test did not, being conservative when sample sizes and variances were directly paired and liberal when the pairing was inverse. When sampling was from skewed distributions, only the Brown-Forsythe test was robust to violations of homogeneity of variance. The ANOVA $\underline{F}$ test and the Welch APDF test were especially prone to inflated Type I error rates when unequal variances were inversely paired with unequal sample sizes. All three tests shared similar power rates when sample sizes were equal and when unequal samples sizes were directly paired with population variances. Clinch and Keselman did not make power comparisons in the case of indirect pairings because only the Brown-Forsythe test adequately controlled Type I error rates.

Tomarken and Serlin (1986) investigated five tests – the ANOVA $\underline{F}$, the Welch APDF, the Brown-Forsythe, and two nonparametric tests – under the following conditions: (a) samples were obtained from normal distributions, (b) $\underline{k} = 3$ or 4, (c) largest to smallest population variance ratio $\theta = 1$, 6, or 12, (d) largest to smallest sample size ratio $\underline{n}_r = 1$ or 3, and (e) $36 \leq \underline{N} \leq 80$. Four configurations of mean, sample size, and variance were used to assess power, using the .01 and .05 levels of significance. Tomarken and Serlin found all three tests to perform acceptably under homoscedastic conditions. When both population variances and sample sizes were equal, the ANOVA $\underline{F}$ was liberal, but for the

Brown-Forsythe and Welch APDF tests actual and nominal Type I error rates closely agreed, the agreement being slightly better for the Welch test. In the positive condition the ANOVA F test was extremely conservative, the Brown-Forsythe test was slightly liberal, though tolerable, and the Welch APDF test showed good control over Type I error rate. In the negative condition the ANOVA F test was extremely liberal, while the Brown-Forsythe and Welch APDF tests were slightly liberal. When variances were equal, the ANOVA F test was the most powerful, followed closely by the Brown-Forsythe test. Under homoscedasticity power level differences were slight. Only the Brown-Forsythe and Welch APDF were compared for power. The Welch APDF test was optimal when means were equally spaced, when one extreme mean was paired with the smallest variance, and when two equal means were halfway between two extreme means. The Brown-Forsythe was optimal only when one extreme mean was paired with the largest variance.

Wilcox, Charlin, and Thompson (1986) observed the behavior of the ANOVA F, Welch APDF, and Brown-Forsythe tests under conditions that extended those of the Brown and Forsythe (1974) study: (a) sampled populations were normal, (b) $k = 2$, 4, or 6, (c) largest to smallest population variance ratio $\theta = 1$ or 4, (d) largest to smallest sample size $n_r = 1$, 1.9, 3, 3.3, or 4.2, (e) $22 \leq N \leq 200$. To assess power the first mean in each group of means was set equal to 1.2; the others were zero. Wilcox, Charlin, and Thompson concluded that the F test was even more sensitive to violations of homoscedasticity than had been previously thought. Even for equal samples of size 50 each the test was not robust for $\theta = 4$ and $k = 4$. The authors agreed with the findings of Brown and Forsythe for the cases they considered ($\theta = 3$), but discovered that neither the Brown-Forsythe test nor the Welch APDF test was robust to heteroscedasticity when sample sizes were unequal and the ratio of the largest to smallest variance was four. They found little loss in power when either the Brown-Forsythe test or the Welch APDF test was used in place of the ANOVA $\underline{F}$ test when variances were not equal. Under heteroscedasticity the Brown-Forsythe test and the Welch APDF test differed drastically in power. The Welch APDF test was usually more powerful, but the reverse was sometimes true. Wilcox, Charlin, and Thompson recommended abandoning the ANOVA $\underline{F}$ test and using the Brown-Forsythe test when variances are homogeneous, especially if sample sizes differ. They recommended using the Welch APDF test when variances differ, but sample sizes do not. None of the tests studied were recommended when sample sizes differ and heterogeneity is extreme ($\theta \geq 4$).

Wilcox (1988) compared his newly proposed $\underline{H}$ statistic with the Welch $\underline{F}_v$, the Brown-Forsythe $\underline{F}^*$, and the James second-order statistic. Conditions reported included: (a) samples were derived from populations that were normal, light-tailed, symmetric, medium-tailed, asymmetric, or exponential-like, (b) $\underline{k}$ = 4, 6, or 10, (c) largest to smallest population variance ratio θ = 1, 4, 5, 6, or 9, (d) largest to smallest sample size ratio $\underline{n}_r$ = 1, 5, 2.5, 3.3, or 1.8, and (e) $44 \leq \underline{N} \leq 106$. For power comparisons the mean of the first group was set to 1.2, 0.4, or 2.4, and the groups were tested under 17 variance and sample size conditions. Wilcox found his proposed $\underline{H}$ test to compare favorably with the James second-order method, giving excellent results with equal sample sizes and revealing a slight liberal tendency when sample sizes differed. Both tests outperformed competitors $\underline{F}_v$ and $\underline{F}^*$. Deviations from normality had little effect on either the Wilcox $\underline{H}$ or the James second-order test. One exception was noted. The Wilcox $\underline{H}$ test was conservative and the James second-order test was liberal under extreme non-normality with moderately small sample sizes. The James second-order test was slightly less powerful than the Welch APDF test and generally more powerful than the Wilcox $\underline{H}$ test except when the largest mean was paired with the smallest variance and smallest sample size. In that case the James second-order test was considerably more powerful.

Wilcox recommended the James second-order test over the $\underline{H}$ test, but proposed a modified $\underline{H}$ statistics, $\underline{H}_m$, designed to compare more favorably in power with the James second-order test (Wilcox, 1989). A new study to compare the James second-order method with the modified Wilcox statistic used: (a) samples from normal populations, (b) $\underline{k} = 4$ or 6, (c) largest to smallest population variance ratio $\theta = 1$, 4, or 6, (d) largest to smallest sample size ratio $\underline{n}_r = 1$, 2.5, 2.7, or 5, and (e) $44 \leq \underline{N} \leq 121$. For power assessment three alternatives were used; the first mean was increased by 1, 2, or 3. Wilcox's results showed the James second-order test to be slightly liberal, while the Wilcox $\underline{H}_m$ test was slightly conservative. The James second-order test was more powerful, but not substantially so. The $\underline{H}_m$ statistic was a clear improvement over $\underline{H}$, but it deteriorated under the same conditions that caused the earlier version to be unsatisfactory – small sample sizes, large numbers of groups, and increased variance ratios.

Hsiung, Olejnik, and Huberty (1994) studied the Wilcox $\underline{H}_m$ test under a wider range of conditions and found the test to be invalid under small, but reasonable, unequal sample sizes and a common population mean different from zero. They further showed the test not invariant to the distribution location parameter, thus, effectively ruling out the test for consideration with interval data, which comprises a large bulk of the data in psychology and education.

Oshima and Algina (1992) included the Welch APDF test, Brown-Forsythe test, James second-order method, and two Wilcox tests ($\underline{H}$ and $\underline{H}_m$) in a Monte

Carlo study that crossed the 31 Wilcox (1988) study conditions with 5 distributions – normal, uniform, $\underline{t}$ with 5 degrees of freedom, beta with parameters 1.5 and 8.5, and exponential. Oshima and Algina did not make power comparisons. They found no single test to be uniformly superior in controlling Type I error rate. In general, the James second-order test was superior to both the Welch APDF test and the Brown-Forsythe test. The modified Wilcox test was superior to the Brown-Forsythe test. The $\underline{H}_m$ test was conservative with normal distributions or long-tailed symmetry; the James second-order test was not. For short-tailed symmetry the James second-order test tended to be more liberal than the Wilcox $\underline{H}_m$ test. Both tests were liberal with asymmetry, the Wilcox test less so. Oshima and Algina conjectured based on these findings that the James second-order test has a power advantage over the modified Wilcox test. They recommended the James second-order test when data are symmetric in distribution and Wilcox $\underline{H}_m$ test with moderate skewness.

A meta-analysis conducted by Harwell, Rubinstein, Hayes, and Olds (1992) summarized the ANOVA $\underline{F}$ and Welch APDF $\underline{F}_V$ tests in terms of both Type I error control and power under assumption violations. When sample sizes are equal, heteroscedasticity has a modest inflationary effect on Type I error rate for $\underline{F}$ that increases as the variance ratio θ increases. The effect on Type I error rate for the Welch $\underline{F}_V$ remains modest for variance ratios as high as 8:1. The effect on power for both tests is negligible. When sample sizes differ, the Type I error rate for the ANOVA $\underline{F}$ is seriously affected by heteroscedasticity, while the Welch $\underline{F}_V$ shows only a slight liberal tendency. The effect on the power of the $\underline{F}$ test is negligible, while $\underline{F}_V$ experiences a slight inflationary trend.

Wilcox (1993) designed a Monte Carlo experiment to assess his $\underline{Z}$ test, a generalization of the Wilcox (1992) $\underline{H}_M$ test. Both were developed to accommodate heavy-tailed distributions which affect power (Yuen, 1974) and have been shown to be common (Micceri, 1989). Wilcox examined $\underline{Z}$ under the following conditions: (a) samples were selected from normal, exponential, uniform, $\underline{t}$ with 5 degrees of freedom, and moderately skewed distributions, (b) $\underline{k} = 4$, (c) largest to smallest population variance ratio $\theta = 1$ or 4, (d) samples sizes were 21 or 41, yielding a ratio of largest to smallest sample size $\underline{n}_r = 1$ or 1.95, and (e) $105 \leq \underline{N} \leq 125$. Results showed that $\underline{Z}$ can be unsatisfactory at the .01 significance level in controlling Type I error rate. At $\alpha = .05$ and .10 control was adequate except when $\underline{Z}$ was used with the relatively light-tailed exponential distribution.

Alexander and Govern (1994) compared the Alexander-Govern $\underline{A}$ statistic with the ANOVA $\underline{F}$ and James second-order statistics. A large range of conditions that threaten control of Type I error rate (large numbers of groups and small sample sizes) were considered in the robustness phase of the study. The power phase examined nine conditions, a result of crossing three effect sizes with three mean patterns. Effect size was defined as range of means and was set at 0.5, 1.0, or 1.5. The three mean patterns used were those of Cohen (1988): maximum variation (half the means at each extreme of the range), minimum variation (one mean at each extreme and the others at the median), and intermediate variation (equally spaced means). In terms of control of Type I error, $\underline{A}$ was found to be similar to the James second-order statistic across all conditions, deviating less than .007 at $\alpha = .05$ and less than .005 at $\alpha = .01$. The power study showed that conditions that inflated Type I error rate for the ANOVA $\underline{F}$ resulted in its power's being lower than that of either $\underline{A}$ or the James statistic. The reverse was also true. The $\underline{A}$ statistic had power levels comparable to those of the James second-order statistic.

On balance, the alternatives to the ANOVA $\underline{F}$ test offer improved control over Type I error rate under heteroscedastic conditions in normal distributions. The James second-order method and the Govern-Alexander $\underline{A}$ test offer the best control, especially as heteroscedasticity increases. The two tests have similar powers. Despite its difficult computations, the James second-order solution appears to be a statistic of choice, along with that of Alexander and Govern, for testing for mean differences under heteroscedastic conditions now that computer code has been written for the test (Oshima & Algina, 1992).

Hotelling's T² Test

Hotelling's (1931) $\underline{T}^2$ test is used to test two population mean vectors of order $\underline{p} \times 1$ for equality under the assumptions that independent samples of sizes $\underline{n}_1$ and $\underline{n}_2$ are randomly selected from two multivariate normal populations with equal covariance matrices, Σ_1 and Σ_2. The test statistic is

$$\frac{n_1 n_2}{n_1 + n_2} (\bar{x}_1 - \bar{x}_2)' \, S^{-1} (\bar{x}_1 - \bar{x}_2) \; ,$$

where

$$S = \frac{(n_1 - 1)S_2 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \; .$$

S_1 and S_2 are the sample covariance matrices. Hotelling used

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

to transform the distribution of $\underline{T}^2$ to an $\underline{F}$ distribution with $\underline{p}$ and $\underline{n}_1 + \underline{n}_2 - \underline{p} - 1$ degrees of freedom. Hsu (1938b) developed the power function of $\underline{T}^2$ and discussed its optimum properties, showing $\underline{T}^2$ to be the most powerful statistic in its class according to the Neyman-Pearson theory of testing of statistical hypotheses.

The behavior of $\underline{T}^2$ when the assumptions of multivariate normality and homoscedasticity are violated have been well documented in the literature. Both analytic (Ito & Schull, 1964; Mardia, 1971; Pillai & Sudjana, 1975) and empirical (Algina & Oshima, 1990; Everitt, 1979; Hakstian, Roed, & Lind, 1979; Holloway & Dunn, 1967; Hopkins & Clay, 1963; Mardia, 1971) studies have been performed.

Hopkins and Clay (1963) studied Hotelling's $\underline{T}^2$ test using Monte Carlo experiments when (a) samples were from bivariate ($\underline{p} = 2$) normal or bivariate symmetric leptokurtic distributions and (b) sample sizes (n_1, n_2) were (5, 5), (5, 10), (10, 5), (10, 10), (10, 20), (20, 10), or (20, 20). Heteroscedasticity was simulated in the bivariate populations with covariance matrices $\Sigma_i = \sigma_i^2 \mathbf{I}$ (i = 1, 2) where $\sigma_2:\sigma_1 = 1$, 1.6, or 3.2. The bivariate symmetrical leptokurtic populations were distributed with zero means, equal variances, and $\beta_2 - 3 = 3.2$ or 6.0. Hopkins and Clay found Hotelling's $\underline{T}^2$ test to be rather robust in terms of Type I error control to violations of the equal covariance matrices assumption for $\underline{n}_1 = \underline{n}_2 > 10$. For unequal sample sizes the authors reported the test to be conservative in the positive condition ($\tau = .01$, $\alpha = .05$, $\sigma_2:\sigma_1 = 3.2$, $\underline{n}_1 = 10$, $\underline{n}_2 = 20$) and extremely liberal in the negative condition ($\tau = .214$, $\alpha = .05$, $\sigma_2:\sigma_1 = 3.2$, $\underline{n}_1 = 20$, $\underline{n}_2 = 10$). The obtained values of τ agreed with those obtained from Hsu's (1938a) analytically deduced formulas ($\tau = .05$ and .23 for the same conditions). Finally, this study suggested that for $\underline{n}_i \geq 10$ ($\underline{i} = 1, 2$) leptokurtosis exerted little effect on actual Type I error rate.

Ito and Schull (1964) examined analytically both control of Type I error rate and power of the Hotelling's $\underline{T}^2$ test for large samples, unequal covariance matrices, and $\underline{p} = 1, 2, 3$, or 4. They found $\underline{T}^2$ to be well behaved in terms of both control of Type I error rate and power in the face of even large differences in covariance matrices for equal-sized and very large random samples. For samples nearly equal in size $\underline{T}^2$ retained robustness under moderate heteroscedasticity. Markedly different sample sizes, however, led to pronounced effects on both actual Type I error rate τ and power by even moderate differences in Σ_1 and Σ_2. For fixed values of $\underline{n}_r = \underline{n}_1$: $\underline{n}_2$ and equal eigenvalues of $\Sigma_1\Sigma_2^{-1}$, actual Type I error rate exceeded the nominal rate ($\tau > \alpha$) when the eigenvalues were less than 1, and the actual rate was exceeded by the nominal rate ($\tau < \alpha$) when the eigenvalues exceeded 1. This tendency for the test to be liberal for $\underline{n}_r > 1$ and equal eigenvalues less than 1 (the negative condition) and conservative for $\underline{n}_r > 1$ and equal eigenvalues greater than 1 (the positive condition) increased with both $\underline{n}_r$ and $\underline{p}$. The power of Hotelling's $\underline{T}^2$ test under heteroscedastic conditions exceeded that of the test when assumptions were satisfied for $\underline{n}_r > 1$ and equal eigenvalues less than one. The opposite result occurred for $\underline{n}_r > 1$ and equal eigenvalues greater than one. No tendencies in the behavior of the power function were found as a function of $\underline{p}$.

Holloway and Dunn (1967) examined Hotelling's $\underline{T}^2$ test for both Type I error rate and power in a Monte Carlo study in which (a) samples were selected from multivariate normal distributions, (b) the number of dependent variables $\underline{p}$ = 1, 2, 3, 5, 7, or 10, (c) $10 \leq \underline{N} \leq 200$ where total sample size $\underline{N} = \underline{n}_1 + \underline{n}_2$, (d) $\frac{\underline{n}_1}{\underline{N}}$ = .3, .4, .5, .6, or .7, and (e) the eigenvalues of $\Sigma_1\Sigma_2^{-1}$ equaled 3 or 10. The study confirmed that Hotelling's $\underline{T}^2$ test (as well as the independent samples $\underline{t}$ test) is robust to violations of equal covariance matrices (homoscedasticity) in terms of controlling for Type I error rate provided samples are of equal size. For fixed heteroscedasticity Hotelling's $\underline{T}^2$ test was found to be conservative in the positive condition and liberal in the negative condition. Actual Type I error rate

τ tended to increase as $\underline{p}$ or $\frac{\underline{n}_1}{\underline{N}}$ increased or as $\underline{N}$ decreased. In terms of power Holloway and Dunn discovered that under heteroscedastic conditions for fixed N, moving the sample sizes closer to one another reduced not only $|\tau - \alpha|$, the difference between actual and nominal Type I error rate, but also power. The effect size required for reasonable power declined as $\underline{N}$ increased. Power was found to be related to $\underline{p}$. For $\Sigma_1 = \Sigma_2$ power declined as $\underline{p}$ increased. Under heteroscedasticity power declined as $\underline{p}$ increased when the ratio between first sample size and total sample size $\left(\frac{\underline{n}_1}{\underline{N}}\right)$ was sufficiently small. For larger values of $\frac{\underline{n}_1}{\underline{N}}$, increasing $\underline{p}$ increased τ, resulting in higher than expected power for small effects and lower than expected power for large effects. The authors concluded that when $\Sigma_1 \neq \Sigma_2$, τ may differ greatly from α, and if τ is too large, power for small effects will be too high, while power for large effects will be too small. These tendencies increased with $\underline{p}$ and with the degree of heteroscedasticity, but decreased with $\underline{N}$. Equal samples sizes were shown to maintain significance level, but were of little or no help in maintaining power under unequal covariance matrices. Hakstian, Roed, and Lind (1979) criticized the Holloway-Dunn study for having unrealistically different covariance matrices (variances in the second population were as much as 10 or 100 times as large as those in the first population) and unequal sample sizes that were not disparate enough (15:35 $\leq$ $\underline{n}_1$: $\underline{n}_2$ $\leq$ 35:15 when $\underline{n}_1 + \underline{n}_2 = 50$).

Pillai and Sudjana (1975) explored mathematically the behavior of four multivariate statistics that reduce to Hotelling's $\underline{T}^2$ test in the two-sample case. Using equal-sized samples of 5, 15, and 40, Pillai and Sudjana reported modest departures from α for even minor heteroscedasticity, departures that became more pronounced as the degree of heteroscedasticity increased.

Everitt (1979), Ito (1969), and Mardia (1971, 1975), all considered

Hotelling's $\underline{T}^2$ test in terms of Type I error control under departures from multivariate normality. In analytic studies Ito (1969) found the test to be robust for very large samples; Mardia (1971) obtained similar results in an empirical study using moderately large samples. Mardia also concluded that for small samples $\underline{T}^2$ is generally robust to non-normality, but shows some sensitivity to skewness when sample sizes differ. Everitt (1979) employed empirical techniques to examine both one- and two-sample $\underline{T}^2$ tests. He examined tests applied to bivariate normal, uniform, exponential, gamma, and lognormal distributions. In the one-sample case highly skewed distributions were not well approximated by $\underline{T}^2$. In the two-sample case Hotelling's $\underline{T}^2$ test was judged fairly robust to non-normality, although departures due to skewness led to moderately, or in some cases extremely, conservative tests. The number of variates $\underline{p}$ and equality of sample sizes appeared to have little effect on actual Type I error rate. Mardia (1975) used a test of multivariate normality based on measures of skewness and kurtosis to interpret Monte Carlo results of various studies.

Hakstian, Roed, and Lind (1979) designed a Monte Carlo study to examine simultaneously all individual variables relative to robustness using conditions that represent real-world behavioral data. Those conditions included: (a) samples selected randomly from multivariate normal populations, (b) $\underline{p} = 2, 6,$ or 10, (c) average number of subjects per variable equal to 3 or 10, (d) sample size ratios $\underline{n}_1$: $\underline{n}_2 = 1, 2,$ or 5. Heteroscedasticity was modeled using covariance matrices of the form **I** and **D** where **D** equaled **I**, $\underline{d}^2$**I**, or diag$\{1, 1, \ldots, 1, \underline{d}^2, \underline{d}^2, \ldots, \underline{d}^2\}$ ($\underline{d} =$ 1, 1.2, or 1.5). Both positive and negative conditions were considered. Hakstian, Roed, and Lind found Hotelling's $\underline{T}^2$ test to be robust to unequal covariance matrices when $\underline{n}_1 = \underline{n}_2$ even for $\frac{\underline{p}}{\underline{N}}$ as small as 3. The robustness did not extend to unequal sample sizes, yielding conservative tests in the positive condition and liberal tests in the negative condition. The authors concluded that Hotelling's $\underline{T}^2$ test is not robust in terms of Type I error control in the face of even mild departures from equal covariance matrices when sample sizes differ. The difference between actual and nominal rate $|\tau - \alpha|$ increased as $\frac{\underline{n}_1}{\underline{n}_2}$ departed from 1 or as $\underline{p}$ increased, but was independent of total sample size $\underline{N}$.

Algina and Oshima (1990) extended previous studies examining the ability of Hotelling's $\underline{T}^2$ to control Type I error rate in the face of unequal covariance matrices by considering small sample size ratios. Their investigation was performed under the following conditions: (a) samples were obtained from multivariate normal distributions, (b) $\underline{p} = 2$, 6, or 10, (c) $\underline{n}_1$: $\underline{n}_2$ = 1:1.25, 1.25:1, 1:1.1, 1.1:1, or approximately 1:1, (d) $\frac{N}{\underline{p}} = 6$, 10, or 20, and (e) for most conditions $\Sigma_2 = d^2\Sigma_1$ (d = 1.5, 2.0, 2.5, or 3.0). Algina and Oshima found Hotelling's $\underline{T}^2$ test to be seriously nonrobust in terms of controlling Type I error rate when $\Sigma_1 \neq \Sigma_2$, even for equal sample sizes, especially if the ratio of total sample size to number of variables was small. They recommended using $\underline{T}^2$ in the positive condition, but suggested that alternatives not sensitive to differences in covariance matrices be considered for the negative condition.

In summary Hotelling's $\underline{T}^2$ test is the most powerful in its class according to Neyman-Pearson theory (Hsu, 1938b). In terms of Type I error control it is fairly robust to non-normality except departures owing to skewness. In that case the test is conservative. The test is robust to unequal covariance matrices as long as sample sizes are equal and the number of subjects per variable $\left(\frac{N}{\underline{p}}\right)$ is large. This robustness, however, fails when the $\frac{N}{\underline{p}}$ ratio becomes small and does not extend to unequal sample sizes. In the positive condition, when the larger sample is selected from the population with the larger dispersion, Hotelling's $\underline{T}^2$ test is conservative; in the negative condition, when the larger sample is selected from the population with the smaller dispersion, the test is liberal. Therefore, numerous situations arise in which alternative to Hotelling's $\underline{T}^2$ test merit consideration.

Alternatives to the Hotelling's $\underline{T}^2$ Test

Various tests have been devised to test the equality of two mean vectors in the presence of assumption violations. Tiku, Gill, and Balakrishnan (1989) and Nath and Duran (1983) proposed tests to operate when the assumption of multivariate normality may be untenable. Tiku, Gill, and Balakrishnan extended their univariate test based on modified maximum likelihood estimators. Nath and Duran proposed applying Hotelling's $\underline{T}^2$ to the rankings of data rather than to the data themselves. While technically not a parametric test, this easy-to-use alternative does, at best, occupy a position that serves as a bridge between parametric and nonparametric procedures. As with univariate data, nonparametric tests have been proposed to handle departures from normal conditions. But multivariate nonparametric techniques, unlike their univariate counterparts, involve computationally complex techniques and their null distributions often require enormous calculations (Nath & Duran, 1983).

Bennett (1951), Andersen (1958), and Ito (1969) extended the work of Scheffé (1943), which in the univariate case produces a technique to deal with the Behrens-Fisher problem that yields the shortest confidence interval using the t distribution (Anderson, 1958). The Scheffé technique of randomization necessitates randomized pairings of observations in different groups and the discarding of data when sample sizes differ (Algina & Tang, 1988).

At least five solutions to the two-sample multivariate Behrens-Fisher problem are based on the same statistic and differ only in their critical values. These five tests which do not assume the two population covariance matrices are equal are the Yao (1965) test, James (1954) first- and second-order tests, Johansen

(1980) test, and a test developed by Nel and van der Merwe (1986). Their test statistic is

$$T_v^2 = (\bar{x}_1 - \bar{x}_2)'\Big(\frac{S_1}{n_1} + \frac{S_1}{n_1}\Big)(\bar{x}_1 + \bar{x}_2)$$

where $\bar{x}_i$ and S_i are, respectively, the sample mean vector and sample covariance matrix for the ith sample ($i = 1, 2$). This test statistic is a multivariate extension of the Welch APDF (1947) statistic. Kim (1993) proposed a test based on the same statistic with A^{-1} substituted for $\Big(\frac{S_1}{n_1} + \frac{S_1}{n_1}\Big)$, where

$$A = A_1 + r^2 A_2 + 2rA_2^{1/2}(A_2^{-1/2}A_1A_2^{-1/2})^{1/2}A_2^{1/2}$$

$$r = |\ A_1A_2^{-1}\ |^{\frac{1}{2p}}$$

and

$$A_i = \frac{S_i}{n_i}\ .$$

The literature suggests the following regarding the Hotelling T^2, Bennett, Yao, Kim, James first- and second-order, and Johansen tests: (a) the alternatives to Hotelling's T^2 test are superior to it in control of Type I error rate when the normality assumption is satisfied, but homoscedasticity is not, (b) the Yao, Kim, James second-order, and Johansen tests are superior to the James first-order test in controlling Type I error rate, (c) all the alternatives to Hotelling's T^2 test are sensitive to skewness in the sampled population, and (d) the James first-order, Kim, and Yao tests have similar powers.

Yao (1965) compared the James first-order test with Yao's test under the following conditions: (a) samples were taken from multivariate normal distributions, (b) $\underline{p} = 2$, (c) $\frac{N}{\underline{p}} = 9$ or 12 when $\underline{n}_1 \neq \underline{n}_2$ ($\underline{n}_1 = 6$, $\underline{n}_2 = 12$ or 18) and $\frac{N}{\underline{p}} = 13$ when $\underline{n}_1 = \underline{n}_2 = 13$, and (d) $\Sigma_1 \neq \Sigma_2$. Yao found the actual Type I error rate τ approximately equal to the nominal rate α for both tests with Yao's test being slightly superior to James's first-order test.

Subrahmaniam and Subrahmaniam (1973) examined empirically the Yao, Bennett, and James first-order tests using: (a) samples from multivariate normal distributions, (b) $\underline{p} = 2, 4, 5$, or 10, (c) $\frac{N}{\underline{p}}$ ranging from 3 ($\underline{n}_1 = \underline{n}_2 = 15$ and $\underline{p} = 10$) to 12 ($\underline{n}_1 = 6$, $\underline{n}_2 = 18$, $p = 2$), and (d) unequal covariance matrices. They found the Yao test to be more conservative than the James first-order test, the James first-order test to be notably inferior in the negative condition, and the James first-order test to deteriorate in Type I error control as $\underline{p}$ increased. Neither test provided the sought-after control. Bennett's test controls Type I error rate exactly, so was not included in the significance level results. In terms of power the James first-order test was found superior, followed closely by the Yao test. The high power levels of these two tests, however, were a reflection of high rates of Type I errors. Bennett's test had poor power. The power levels of all three tests declined as the number or dependent variables increased.

Algina and Tang (1988) conducted Monte Carlo experiments extending Yao's study which considered only the behavior of bivariate data. Algina and Tang's work included the following conditions: (a) samples were drawn from multivariate normal populations, (b) $\underline{p} = 2, 6$, or 10, (c) $\underline{n}_1$: $\underline{n}_2$ = 1:5, 1:4, 1:3, 1:2, 1:1.5, 1:1.25, 1, 1.25, 2, 3, 4, or 5, (e) covariance matrices were of the form I and D where $D = \underline{d}^2 I$ ($\underline{d} = 1.5, 2.0, 2.5$, or 3.0) or of the form I and D where D was diag$\{3, 1, 1, . . . 1\}$, diag$\{3, 3, . . . ,3, 1, 1, . . . , 1\}$, diag$\{\frac{1}{3}, 3, 3, . . . , 3\}$, or diag$\{\frac{1}{3}, \frac{1}{3}, . . . ,\frac{1}{3}, 3, 3, . . ., 3\}$. Algina and Tang confirmed the superiority of the Yao test over the James first-order test and the James first-order test over Hotelling's

$\underline{T}^2$ test in all studied conditions. Noting their conclusions were limited by the range of values studied for $\underline{p}$, $\underline{d}$, $\underline{n}_1$: $\underline{n}_2$, $\frac{N}{\underline{p}}$, and the degree of difference between the covariance matrices, Algina and Tang judged Yao's test safe as far as controlling Type I error rate for equal sample sizes when $10 \leq \frac{N}{\underline{p}} \leq 20$. The test became liberal for $\underline{n}_1 = \underline{n}_2$ when $\frac{N}{\underline{p}} = 6$, $p \geq 10$, and $d \geq 3.0$. For differing sample sizes, Algina and Tang concluded that Yao's test can be safely used provided $\underline{p} \leq 10$, $\frac{N}{\underline{p}} \geq 10$, and the ratio of the larger to smaller sample size is 2:1 or smaller. If $\frac{N}{\underline{p}}$ exceeds 20, Yao's test can be safely used for $\underline{p} = 2$ and a sample size ratio of 5 or less, $\underline{p} = 6$ and a sample size ratio of 3 or less, and $\underline{p} = 10$ and a sample size ratio of 4 or less.

Algina, Oshima, and Tang (1991) studied the Yao, James first- and second-order, and Johansen tests under various combinations of heteroscedastic conditions and departures from normality. Sampled distributions were normal, uniform, Laplace, t with 5 degrees of freedom, beta (5, 1.5), exponential, and lognormal. The conditions considered were those recommended as safe for Yao's test under multivariate normality by Algina and Tang (1988). Tests were studied in both the positive and negative conditions. For all alternatives to Hotelling's $\underline{T}^2$ test considered, actual Type I error rate τ was in the interval [.025, .075], Bradley's (1978) liberal criterion for robustness at the .05 significance level. Asymmetry resulted in elevated significance levels, the degree of elevation depending on the degree of asymmetry, the degree and pattern of heteroscedasticity, the ratio of the largest to smallest sample size, and the number of dependent variables. For moderate asymmetry (beta distribution), τ tended to be in the robust interval even for large departures from homoscedasticity and large sample size ratios. For small degrees of heteroscedasticity and a sample ratio of $\underline{n}_1$: $\underline{n}_2$ = 1.5:1, the tests were robust even for the extremely asymmetric lognormal distribution. The James first-order test was slightly inferior to the other three alternatives, none of which had a clear advantage in terms of significance level.

Kim (1992) compared the Kim and Yao tests for both Type I error control and power under the following conditions: (a) samples were selected from multivariate normal populations, (b) $p = 2$ or 5, (c) $\frac{N}{p}$ ranged from 3.6 ($n_1 = 6$, $n_2 = 12$, $p = 5$) to 16 ($n_1 = 8$, $n_2 = 24$, $p = 2$), (d) covariance matrices were of the form diag$\{\frac{1}{9}, \frac{1}{9}\}$, diag$\{9, 9\}$, diag$\{\frac{1}{9}, 9\}$, diag$\{5, 9\}$, diag$\{\frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}, \frac{1}{9}\}$, diag$\{9, 9, 9, 9, 9\}$, diag$\{\frac{1}{9}, \frac{1}{5}, 1, 5, 9\}$, and diag$\{5, 6, 7, 8, 9\}$. The Yao and Kim tests were found to be very similar in both Type I error control and power, although Yao's test did have inflated rates as high as .172 in the negative condition. Yao's test showed slight power advantages in the positive condition; whereas, Kim's test had better power in the negative condition.

In summary, all of the alternatives considered to Hotelling's T^2 test perform well under heteroscedastic conditions. James first-order test is inferior to the Kim, Yao, James second-order, and Johansen tests. All have larger than nominal Type I error rates when sampling is from skewed distributions. Since the Yao, Kim, James second-order, and Johansen tests are comparable in terms of significance level, distinctions must be made based on practical considerations such as ease of use, availability of computer code, and power. At present Johansen's test enjoys a slight advantage.

Classic Multivariate Analysis of Variance

Four classic criteria are used to test for the equality of k population mean vectors when independent samples are selected from populations distributed multivariate normally (p dependent variables) with equal covariance matrices. These four multivariate analysis of variance (MANOVA) criteria are all defined in terms of the eigenvalues of HE^{-1} where

$$H = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$

and

$$E = \sum_{i=1}^{k} (n_i - 1)S_i \quad .$$

The matrix H is the hypothesis sum of squares and cross products matrix, and E is the error sum of squares and cross products matrix. If λ_i is the ith eigenvalue of HE^{-1} ($i = 1, 2, \ldots s$), where $s = \min(p, k-1)$, the four criteria are:

1. Roy's (1945) largest root criterion
$$R = \frac{\lambda_1}{1 + \lambda_1}$$
where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s$;

2. Hotelling-Lawley trace criterion (Hotelling, 1951; Lawley, 1938)
$$U = trace(HE^{-1}) = \sum_{i=1}^{s} \lambda_i \quad ;$$

3. Wilks's (1932) likelihood ratio criterion
$$L = \frac{|E|}{|H + E|} = \prod_{i=1}^{s} \frac{1}{1 + \lambda_i} \quad ;$$

4. Pillai-Bartlett trace criterion (Bartlett, 1939; Pillai, 1955)
$$V = trace[H(H + E)^{-1}] = \sum_{i=1}^{s} \frac{\lambda_i}{1 + \lambda_i} \quad .$$

The statistics U, L, and V are asymptotically equivalent in very large samples. Olson (1976) noted that many multivariate tests reported in the literature fail to identify the criterion used regardless of sample size. He recommended not only reporting the statistic, but also specifying the approximation, if any, used. Numerous approximations and transformations have been derived for the MANOVA criteria. Pillai (1956) provided tables for an approximate distribution of Roy's largest root criterion. Ito (1960), Hughes and

Saw (1972), and Pillai and Samson (1959) approximated the Hotelling-Lawley criterion. McKeon (1974) improved upon the latter two approximations.

Rao (1948), Posten and Bargmann (1964), and Sugiura and Fujikoski (1969) proposed asymptotic formulae for the Wilks likelihood ratio statistic. Lee (1972) examined L in both its exact and asymptotic forms. An asymptotic formula for the Pillai-Bartlett trace criterion was suggested by Lee (1971). According to Olson (1976), Pillai's (1960) F approximation to V is a good one.

Elliott and Barcikowski (1994) examined F approximations for U and V in canned computer packages when assumptions are met. They concluded that with small numbers of subjects (15 or fewer per group) SAS(GLM) and SPSS(MANOVA) are conservative for V and liberal for U. BMDP4V was found to be accurate for both. All three programs were accurate for L and R. The power of the U and V criteria using F approximations was very near that computed using critical values found through Monte Carlo techniques.

Robustness tests for the MANOVA criteria have been both analytic (Ito & Schull, 1964; Pillai & Sudjana, 1975) and empirical (Korin, 1972; Olson, 1974).

Ito and Schull (1964) limited their analytic study to the behavior of U, the Hotelling-Lawley likelihood ratio criterion, under conditions of unequal covariance matrices. They concluded that if samples are of equal size, moderate inequality in covariance matrices does not affect Type I error rate seriously as long as samples are very large. For unequal samples of any size, however, quite large effects were observed in the significance level under heteroscedastic conditions.

Korin (1972) conducted a Monte Carlo experiment examining Roy's largest root criterion R, the Hotelling-Lawley trace criterion L, and the Wilks likelihood ratio criterion L under the following conditions: (a) samples were selected from normal populations, (b) samples selected were both equal and unequal in size, (c)

$\underline{k} = 3$ or 6, (d) $\underline{p} = 2$ or 4, (e) $\frac{N}{\underline{p}} = 8.25$, 9, 12, 15.5, 18, or 33, and (f) covariance matrices were {I, I, $\underline{d}$I} or {I, $\underline{d}$I, 2$\underline{d}$I} for $\underline{k} = 3$ and {I, I, I, I, I, dI} or {I, I, I, I, $\underline{d}$I, 2$\underline{d}$I} for $\underline{k} = 6$ ($\underline{d} = 1.5$ or 10). Korin found no great differences among $\underline{R}$, $\underline{U}$, and $\underline{L}$, although $\underline{L}$ was found to be liberal, $\underline{U}$ even more so, and $\underline{R}$ the most liberal. Small heteroscedasticity led to mild departures from nominal Type I error rates, large heteroscedasticity to larger departures.

Pillai and Sudjana (1975) studied all four criteria, but limited their study and, hence, its generalizability to two populations ($\underline{k} = 2$). Results suggested the degree of departure from α increases with the degree of heteroscedasticity, results agreeing with both the Ito and Schull (1964) and the Korin (1972) studies.

Olson (1974) compared the robustness of six multivariate tests based on the eigenvalues of HE^{-1}, including the four classic MANOVA criteria. Conditions were: (a) samples were selected from normal or kurtotic populations, (b) samples were of equal sizes 5, 10, or 50, (c) $\underline{k} = 2$, 3, 6, or 10, (d) $\underline{p} = 2$, 3, 6, or 10, (e) covariance matrices were of the form **I** or **D**, and (f) either high degree of contamination or low degree of contamination was present. High contamination was modeled with $\mathbf{D} = \mathrm{diag}\{\underline{pd} - \underline{p} + 1, 1, 1, 1, \ldots 1\}$ and low contamination was modeled with $\mathbf{D} = \underline{d}\mathrm{I}$ ($\underline{d} = 1$, 4, 9, or 36). To study and assess power Olson used either a concentrated or a diffuse noncentrality structure. In a concentrated noncentrality structure one group differs from all other groups on just one variable or on all variables, but no other differences exist. In a diffuse structure each group differs from all other groups on just one variable, but no other differences exist. Olson's results reinforced those of earlier investigators by showing increased departures from nominal Type I error rates (conservative results became more conservative and liberal results became more liberal) as $\underline{d}$, and hence, heteroscedasticity increased. The Pillai-Bartlett trace criterion $\underline{V}$ responded least to dispersion differences. Whereas *R*, *U*, and *L* experienced increased exceedance rates with increases in the number of dependent variables under low concentration of contamination, *V* did not. Increases in the number of dependent variables under high concentration of contamination resulted in no consistent effects. For protection against kurtosis, *V* was also found to be superior to *R*, *U*, and *L* in terms of controlling Type I error rate. Olson recommended the Pillai-Bartlett trace criterion for general protection against heteroscedasticity and non-normality even though the Hotelling-Lawley trace criterion and the Wilks likelihood ratio criterion are sometimes more powerful. Elliott and Barcikowski (1994) replicated Olson's conditions and agreed with his conclusions that *V* is the most robust of the four classic MANOVA criteria and that it possess suitable power.

Olson (1976) repeated his recommendation to choose the Pillai-Bartlett trace statistic *V* for general protection against both non-normality and unequal covariance matrices. He cited numerous studies (Ito, 1969; Ito & Schull, 1964; Korin, 1972; Mardia, 1971; Olson, 1974) showing that positive kurtosis has mild effects on all four criteria in the conservative direction, the least affected being *V*, followed by *L*, *U*, and *R*, in that order. He reiterated his earlier findings that under heterogeneity of covariance matrices, *R* has excessively high Type I error rates with *U* and *L* close behind. The least severe increases in Type I error rate are reflected in the Pillai-Bartlett trace criterion.

Stevens (1979) took issue with Olson's (1976) recommending *V* for general use. Stevens argued that Olson's claim of *V*'s superiority holds only for extreme subgroup violations, which are rare in practice. He agreed that *V* is best for diffuse noncentrality structures, but contended that *U* or *L* should be used in concentrated noncentrality, since they are slightly more powerful. Olson (1979) responded that noncentrality is a population property and, hence, unknown to the researcher. Since the Pillai-Bartlett trace statistic V is consistently more robust and sometimes more powerful, it is the clear choice when centrality structure is unknown or unclear.

While much has been contributed to the literature in the area of power of multivariate tests when assumptions are met, the subject has not been fully illuminated owing to the vast number of possible conditions and alternatives. No invariant test has the property of uniformly greatest power (Anderson, 1958). Different tests are most powerful in different situations, depending on the nature of the departure from the null hypothesis. Hsu (1940) showed analytically that for large sample sizes the powers of the Hotelling-Lawley trace criterion U and the Wilks likelihood ratio criterion L are equal against all alternatives. Ito (1960) described analytically certain properties of the Hotelling-Lawley trace criterion that help determine power for moderately large samples. Ito (1962) extended the power theory of U and L to samples of moderate size, finding little distinction in power. Gnanadesikan, Lauh, Snyder, and Yao (1965) considered five MANOVA criteria, including three of the classic ones. The nonclassic criteria were found to have power advantages in Monte Carlo studies.

Pillai and Dotson (1969) considered individual roots as test criteria. They concluded the largest, Roy's criteria R, is generally more powerful than other individual root criteria. The authors provided extensive power function tables for two and three dependent variables. Pillai and Jayachandran (1967) considered the four classic criteria in an analytic study with two dependent variables ($p = 2$), when MANOVA assumptions are satisfied. They found V to be most powerful for small deviations from the null hypothesis. The power of Roy's largest root criterion R was less than the powers of the other three criteria. Lee (1971) agreed with the conclusions of Pillai and Jayachandran for $p = 2$, but extended their study to more dependent variables. They discovered that for moderately large samples and small to moderate deviations from the null hypothesis, no single test of $\underline{V}$, $\underline{L}$, and $\underline{U}$ is superior to the other two in terms of power under all alternatives. Stevens (1980) reported that all MANOVA criteria have power problems with small group sizes even for moderate effects. Lauter (1978) created tables that provide the minimum equal sample sizes required for specified powers (.7, .8, .9, and.95) when using the Hotelling-Lawley trace criterion for one of two values of α (.05 or .01), various numbers of dependent variables $\underline{p}$, and three specified alternatives. Cohen (1988) includes tables and examples for determining minimum sample sizes for other MANOVA criteria.

Schatzoff (1966) conducted a Monte Carlo experiment for the purpose of providing data based on ESL (expected significance level, which is equal to $1 - \text{power}$) that may be used as a basis for choosing among the competing criteria $\underline{R}$, $\underline{U}$, $\underline{L}$, and $\underline{V}$. Conditions were (a) samples selected from normal distributions, (b) equal-sized samples, (c) hypothesis degrees of freedom of 2, 4, or 6, (d) $\underline{p} =$ 2, 4, or 6, and (e) diffuse or concentrated noncentrality structures. Schatzoff found the relative power ranking of the four tests was not affected by changes in $\underline{p}$, the number of dependent variables, or $\underline{k}$, the number of groups. Roy's largest root criterion test, however, became increasingly poor relative to the others in terms of power as dimensionality ($\underline{p}$ and/or $\underline{k}$) increased. As sample sizes increased, power increased for all criteria. The powers of $\underline{U}$, $\underline{L}$, and $\underline{V}$ bunched together as sample sizes became very large, not surprising since the three criteria are asymptotically equivalent. Schatzoff found no effect on the ordering of the powers of the classic criteria as the departure from the null hypothesis increased. However, changes in the the noncentrality structure had a large effect on the power rankings. The power of Roy's largest root increased as noncentrality became more concentrated, while the powers of the other criteria eroded. Roy's largest root criterion $\underline{R}$ was the least powerful except when the noncentrality structure was concentrated. Under diffuse noncentrality structures the power rankings were: power($\underline{V}$) > power($\underline{L}$) > power($\underline{U}$) > power($\underline{R}$). The magnitudes of the differences varied inversely with sample size. Schatzoff recommended avoiding Roy's largest root criterion $\underline{R}$ and using either the Hotelling-Lawley trace criterion $\underline{U}$ or the Wilks likelihood ration criterion $\underline{L}$, since both appear to provide good protection against a wide spectrum of alternatives. He argued further against the use of the Pillai-Bartlett trace criterion $\underline{V}$ with small samples and large numbers of dependent variables ($\underline{p}$) or groups ($\underline{k}$), since $\underline{V}$ has low power in situations of concentrated noncentrality, a recommendation in agreement with Stevens (1979), but counter to the suggestion of Olson (1974, 1976, 1979).

To summarize power findings of MANOVA when assumptions are met, $\underline{U}$, $\underline{L}$, and $\underline{V}$ are asymptotically equivalent as sample size increases without bound. Under a concentrated noncentrality structure the power ranking of the four classic criteria is $\underline{R}$, $\underline{U}$, $\underline{L}$, and $\underline{V}$ in order from highest to lowest. When noncentrality structure is diffuse, the power ranking is in the inverse order.

Little has been written about power of the MANOVA criteria under assumption violations. Ito and Schull (1964) concluded that in the case of $\underline{k}$ samples, if the samples are of equal size, moderate inequalities of covariance matrices do not affect the Hotelling-Lawley $\underline{U}$ test as long as samples are very large. But when samples are unequal in size, large effects occur in the test's power. Ito and Schull found no tendency for the power of the test to behave as a function of the number of dependent variables $\underline{p}$. Pillai and Sudjana (1975) studied the four criteria under small levels of heteroscedasticity and found all four to exhibit only modest changes.

Olson (1974) studied the effects of departures from normality and homoscedasticity, which he called contaminations. He found that contamination decreased power, but that these effects were mitigated if there was noncentrality in a group or dependent variable which was contamination free. This was the critical feature to maintain power levels under contamination: the noncentrality had to occur in a noncontaminated group or variable. The power of all four classic tests suffered under kurtosis with the greatest effect occurring for R in a diffuse noncentrality structure. Heteroscedasticity caused all power curves to be rather flat.

To summarize the power of the MANOVA criteria under assumption violations, both kurtosis and heteroscedasticity attenuate power. The effect is important even for small departures and equal sample sizes, and especially when violations occur in contaminated groups or variables. So it is not surprising that alternatives have been sought.

Alternatives to Classic Multivariate Analysis of Variance

Various statistical tests have been devised to address problems encountered with the classic MANOVA procedures when the assumption of equal covariance matrices is untenable. James (1954) generalized the James (1951) series solution to extend to the testing of k mean vectors yielding the statistic

$$J = \sum_{i=1}^{k} (\bar{x}_i - \bar{x})' W_i (\bar{x}_i - \bar{x})'$$

where

$$W_i = \left(\frac{S_i}{n_i}\right)^{-1}, \qquad W = \sum_{i=1}^{k} W_i \, , \qquad \bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \, , \quad \textit{and} \qquad \bar{x} = W^{-1} \sum_{i=1}^{k} W_i \bar{x}_i \, .$$

Zero-, first-, and second-order solutions proposed by James (1954) are analogous to those proposed by James (1951).

Johansen (1980) generalized the Welch (1951) test to the multivariate case resulting in the test statistic

$$C = \frac{\sum_{i=1}^{k} (\bar{x}_i - \bar{x})' W_i (\bar{x}_i - \bar{x})'}{p(k-1) + 2A - \frac{6A}{p(k-1) + 2}}$$

where

$$A = \sum_{i=1}^{k} \frac{trace(I - W^{-1} W_i)^2 + trace^2(I - W^{-1} W_i)}{2(n_i - 1)}$$

and **W**, $\bar{\mathrm{x}}_i$, and $\bar{\mathrm{x}}$ are as defined in the James (1954) statistic. Johansen's statistic follows an F distribution with

$$p(k-1) \qquad \text{and} \qquad p(k-1)\left(\frac{p(k-1) + 2}{3A}\right)$$

degrees of freedom.

Coombs (1993) noted that past researchers (Clinch & Keselman, 1982) had recommended uniform use of the Brown-Forsythe (1974) statistic over the Welch (1951) APDF statistic in the univariate case because of its superior protection against lack of normality. Since the Brown-Forsythe test does not require equal variances, Coombs and Algina (in press) suggested that a multivariate extension of the Brown-Forsythe criterion might produce a test superior to that of Johansen (1980), which is a multivariate generalization of the Welch (1951) test. Coombs and Algina (in press) extended the Brown-Forsythe test by constructing test statistics analogous to the four classic MANOVA criteria and determined approximate degrees of freedom needed to compute critical values using a technique generalized by Nel and van der Merwe (1986) from Satterthwaite's

(1946) univariate method. The resulting four test statistics are

$$R^* = \frac{\lambda_1^*}{1 + \lambda_1^*}$$

$$U^* = trace(HM^{-1}) = \sum_{i=1}^{s} \lambda_i^*$$

$$L^* = \frac{|M|}{|H + M|} = \prod_{i=1}^{s} \frac{1}{1 + \lambda_i^*}$$

and

$$V^* = trace[H(H + M)^{-1}] = \sum_{i=1}^{s} \frac{\lambda_i^*}{1 + \lambda_i^*}$$

where

$$H = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})'$$

and

$$M = \left(\frac{f}{k-1}\right) \sum_{i=1}^{k} (1 - \frac{n_{\mathrm{i}}}{N})S_{\mathrm{i}} \quad .$$

Here, f is the number of degrees of freedom for **M** and ι_{i}^* is the ith eigenvalue of $\mathbf{HM}^{-1}$. Transformations to the F distribution were accomplished using adaptations of the Hughes and Saw (1972) and McKeon (1974) transformations for U, the the Rao (1952) transformation for L, and the Pillai (1960) transformation for V. Because no simple F approximation exists for R, R* was not transformed. Including the two transformations for U*, five tests resulted. They are called the Coombs-Algina R*, the Coombs-Algina U_1^*, the Coombs-Algina U_2^*, the Coombs-Algina L*, and the Coombs-Algina V* tests.

Because the Coombs-Algina tests are all generalizations of the Brown-Forsythe test and because the statistics in those tests are all functions of the same matrix, namely $\mathbf{HM}^{-1}$, similarities in their behaviors under identical experimental conditions should be anticipated. Further, $\underline{U}$, $\underline{L}$, and $\underline{V}$ (the analogs of $\underline{U}_1^*$, $\underline{U}_2^*$, $\underline{L}^*$, and $\underline{V}^*$) are asymptotically equivalent, which should also lead to expectations of similar behaviors in $\underline{U}_1^*$, $\underline{U}_2^*$, $\underline{L}^*$, and $\underline{V}^*$. And the similarities might be expected to increase with total sample size given the asymptotic natures of their analogs. However the Coombs-Algina statistics are not equivalent. While all are functions of $\mathbf{HM}^{-1}$, they are not the same function of $\mathbf{HM}^{-1}$. So, one might emerge as superior under certain conditions, just as the Pillai-Bartlett test has been shown to outperform its competitors, the Wilk's likelihood ratio test and the Hotelling-Lawley trace criterion test, in given experimental situations.

One other attempt to reduce the difference between actual and nominal Type I error rates resulting from heteroscedasticity merits mention. Gabriel (1968) proposed and Bird and Hadzi-Pavlovic (1983) extended simultaneous test procedures (STPs), which are follow-up tests derived from any of the classic MANOVA procedures. The goal was to use power advantages, especially of the STP based upon Roy's largest root test, while at the same time bringing the actual Type I error rate τ near to the nominal Type I error rate α. Tang and Algina (1993) presented a case to show that STPs as alternatives to the Johansen (1980) and James (1954) tests require much more investigation before recommendations regarding their usefulness can be made.

The literature suggests the following in regard to control of Type I error rate when the assumption of equal covariance matrices is violated: (a) of the four classic MANOVA criteria, the Pillai-Bartlett trace criterion $\underline{V}$ is the most robust, (b) even for equal sample sizes, $\underline{V}$ can be liberal, (c) the Johansen test is usually the most robust when sample sizes are equal, (d) the James second-order and Johansen tests are superior to the James zero- and first-order tests and all classic MANOVA criteria when sample sizes differ, (e) when samples are large, Johansen's test is usually most robust, and (f) when samples are small, the James second-order test or the Coombs-Algina $\underline{U}^*$ tests are usually most robust.

In an analytic study, Ito (1969) found the James zero-order test to be liberal. Actual Type I error rate increased with sample size, number of dependent variables, and degree of heteroscedasticity.

Tang and Algina (1993) examined the Pillai-Bartlett trace criterion $\underline{V}$, the Johansen test, and the James first- and second-order tests under the following conditions: (a) samples were selected from normal distributions, (b) $\underline{k} = 3$, (c) $\underline{p} = 3$ or 6, (d) the ratio of the largest to the smallest sample size $\underline{n}_r = 1$, 1.3, or 2, (e) $\frac{\underline{N}}{\underline{p}} = 10$, 15, or 20, and (f) covariance matrices were of the form **I** or **D** where **D** $=$ d**I** or diag$\{1, \underline{d}^2, \underline{d}^2\}$ or diag$\{\frac{1}{\underline{d}^2}, \underline{d}^2, \underline{d}^2\}$ for three dependent variables and **D** $=$ **I** or diag$\{1, 1, 1, \underline{d}^2, \underline{d}^2, \underline{d}^2\}$ or diag$\{\frac{1}{\underline{d}^2}, \frac{1}{\underline{d}^2}, \frac{1}{\underline{d}^2}, \underline{d}^2, \underline{d}^2, \underline{d}^2\}$ for six dependent variables ($\underline{d} = \sqrt{1.5}$ or 3). The authors found that no test studied performed uniformly well. For equal sample sizes, of the tests studied Johansen's test was the most robust in the most conditions to heteroscedasticity. The James first-order solution was liberal and the James second-order solution was conservative. When sample sizes differed, the James second-order and Johansen tests were judged best, although the James second-order test tended to be conservative, while the Johansen test tended to be liberal. James's second-order test was preferred when the ratio of total sample size to number of dependent variables was small.

Coombs and Algina (in press) compared four Coombs-Algina tests, $\underline{U}_1^*$, $\underline{U}_2^*$, $\underline{L}^*$, and $\underline{V}^*$ with the Johansen test under heteroscedastic conditions using unequal sample sizes; specifically: (a) samples were selected from normal or exponential (skewed and kurtotic) distributions, (b) $\underline{k} = 3$ or 6, (c) $\underline{p} = 3$ or 6, (d) the ratio of largest to smallest sample size $n_r = 1.3$ or 2, (e) $\frac{\underline{N}}{\underline{p}} = 10$ or 20, (f) covariance matrices were of the form **I** or **D** with $\mathbf{D} = \text{diag}\{1, d^2, d^2\}$ or $\text{diag}\{1, 1, d^2, d^2, d^2, d^2\}$ ($d = \sqrt{2}$ or 3), and (g) both the positive and negative conditions were explored. All four Coombs-Algina tests outperformed the Johansen test in controlling Type I error rate under the range of conditions considered. Under normality all four Coombs-Algina tests were liberal when $d = 3$ and the condition was positive; otherwise, all were conservative. The L^* test was judged most effective except when $\frac{N}{p}$ was small and k was large, in which case U_1^* provided the best protection against unequal covariance matrices. The Coombs-Algina tests tended to be less sensitive to skewness and kurtosis than the Johansen test. Coombs and Algina (1994) recommended the use of U_1^* with small samples and the use of the Johansen test with large samples when normality holds.

In summary, use of the four classic MANOVA criteria should be avoided when heteroscedasticity is suspected even if sample sizes are equal. Since the prudent researcher always suspects dispersion in the covariance matrices unless there is evidence to the contrary, an argument can be made for the routine use of alternatives not requiring homoscedasticity if they can be shown to maintain Type I error rates and sufficient power levels under a broad range of experimental conditions. Under normality the Coombs-Algina U_1^* test or the James second-order test are most promising for small samples and the Johansen test appears to be the best choice when samples are large. When normality is violated, the Coombs-Algina tests may have advantages. Further investigation involving expanded conditions and power analyses is needed for further illumination.

Summary

Extensive research has examined the Behrens-Fisher problem in the univariate case, both for two populations and k populations. Likewise numerous investigations have dealt with the two-population multivariate case. From these studies have emerged some general themes pointing to the inadequacies of classical tests based on the assumptions of homoscedasticity and normality. Alternative tests not requiring one or both of these assumptions have been developed that have proven effective under a reasonably wide range of conditions in the univariate and two-population multivariate cases. Type I error rate (or robustness) studies have been extensive; power studies have been somewhat limited.

Less has been accomplished in the most general category of two or more multivariate populations. A few landmark studies, most notably Olson (1974), have described the classic MANOVA criteria under assumption violations. Until recently only two alternative tests, not premised on the traditional normality and homoscedasticity assumptions, have been proposed – the James series and Johansen procedures. Only a few researchers have examined the Type I error rates of these tests and no comprehensive power studies have been completed. The proposal of the Coombs-Algina criteria and the results of initial robustness studies suggest that these tests may offer practitioners viable choices in research areas in which heretofore little has been available. But the technical merits of all the MANOVA alternatives have yet to be established. Extensive studies of the James procedures will probably not be undertaken until computer code has been written for the criteria. Technology now makes it possible, however, to delve deeply into the behaviors of both the Johansen and Coombs-Algina criteria. Examinations of complex interactions that provide insight into both Type I error rate and power level are now possible, and the tools by which to accomplish such examinations are readily available.

Chapter 3
Method

This chapter describes the study design and simulation procedure. A Monte Carlo experiment was performed in two phases. The first phase involved the computation of actual Type I error rates for all criteria considered over a range of experimental conditions. Samples were selected from simulated multivariate populations exhibiting violations of underlying test assumptions in which the null hypothesis of equal mean vectors was true. The percentage of time the null hypothesis was rejected (the actual Type I error rate) was computed for each test statistic considered in this study.

In the second phase samples from multivariate populations with the same assumption violations as those used in the first phase were used. The population mean vectors, however, differed. The percentage of time the null hypothesis was rejected (power of the test) was computed for each test statistic considered.

The following sections define the factors that were varied to create assumption violations and other factors that may interact with them.

Design Factors for Robustness Study

The first problem in any robustness study is to choose from the theoretically infinite number of ways in which the assumptions can be violated. This study builds upon and refines the conditions studied by numerous researchers, including Coombs (1993), Tang and Algina (1993), and Olson (1974). Seven factors were varied for the purpose of assessing the robustness of the tests considered to assumption violations in terms of Type I error rate.

Distribution type (DT). Two types of distributions – the normal and exponential – were included in this study. For the normal distribution the coefficients of skewness $\left(\sqrt{\frac{\mu_3^2}{\mu_2^3}}\right)$ and kurtosis $\left(\frac{\mu_4}{\mu_2^2} - 3\right)$ are 0.00 and 0.00, respectively. For the exponential distribution the coefficients of skewness and kurtosis are 2.00 and 6.00, respectively.

A study conducted by Micceri (1989) supports the use of the proposed distributions as reasonable representations of the types of distributions most often encountered in real-world educational and psychological research. Of the 440 distributions examined by Micceri, 60% were the result of published research and 33% came from state, district, or university scoring programs. The results showed that 15.2% of the distributions had both tail weights at or about normality, 18% were less than normality, 17.7 % were moderately higher than that of normality, and 49.1% extremely exceeded normal curve tail weights. Of the 440 distributions, 28.4% were relatively symmetric, 40.7% were moderately asymmetric, and 30.9% were extremely skewed. Of the 30.9% showing extreme skewness, 11.4% were included in a category with skewness coefficients of 2.00 or more.

Number of groups (*k*). Either 3 or 6 populations were sampled in this study. These are the same numbers of populations examined by Korin (1972), Tang (1989), and Coombs (1993). Dijkstra and Werter (1981) used $k = 3$, 4, or 6 and Olson (1974) used simulations involving 2, 3, 6, or 10 groups. Other researchers incorporating either 3 or 6 populations include Brown and Forsythe (1974), Kohr and Games (1974), Clinch and Keselman (1982), Tomarken and Serlin (1986), Wilcox, Charlin, and Thompson (1986), and Wilcox (1988, 1989). The selection of $k = 3$ or 6 appears to be consistent with the literature.

Number of dependent variables (*p*). Data were generated to simulate experiments with a dimensionality of $p = 3$ or 6. The selection of 3 or 6 dependent variables reflects the common range of usage in educational inquiry (Algina & Oshima, 1990; Algina & Tang, 1988; Coombs & Algina, in press; Hakstian, Roed, & Lind, 1979).

Sample size ratio form ($\underline{F}$). Sample size ratio form was modeled after the pattern of Coombs and Algina (in press). As in that study only samples with unequal sizes were selected. The forms of ratios of $\underline{n}_1$: $\underline{n}_2$: $\underline{n}_3$ for $\underline{k} = 3$ and those of $\underline{n}_1$: $\underline{n}_2$: $\underline{n}_3$: $\underline{n}_4$: $\underline{n}_5$: $\underline{n}_6$ for $\underline{k} = 6$ are shown in Table 2. Type I error rate and sample size ratio are positively correlated (Algina & Oshima, 1990). So, if a test performs well with large ratios, it should perform at least equally well with smaller ratios, including equal sample sizes whose ratio is 1. Thus, in the present study sample size ratios considered are as extreme as $\underline{n}_{lg}$: $\underline{n}_{sm}$ = 2:1, where $\underline{n}_{lg}$ is the largest and $\underline{n}_{sm}$ is the smallest sample size.

Table 2

Sample Size Ratio Forms

$\underline{k} = 3$ ($\underline{n}_1$: $\underline{n}_2$: $\underline{n}_3$)

$\underline{n}_1$:	$\underline{n}_2$:	$\underline{n}_3$
1	1	1.5
1	1	2
1	1.5	1.5
1	2	2

$\underline{k} = 6$ ($\underline{n}_1$: $\underline{n}_2$: $\underline{n}_3$: $\underline{n}_4$: $\underline{n}_5$: $\underline{n}_6$)

$\underline{n}_1$:	$\underline{n}_2$:	$\underline{n}_3$:	$\underline{n}_4$:	$\underline{n}_5$:	$\underline{n}_6$
1	1	1	1	1.5	1.5
1	1	1	1	2	2
1	1	1.5	1.5	1.5	1.5
1	1	2	2	2	2

Ratio of smallest sample size to number of dependent variables ($\underline{r} = \frac{\underline{n}_{sm}}{\underline{p}}$). The ratios chosen for simulation were $\underline{r} = 2$, $\underline{r} = 4$, and $\underline{r} = 6$. These ratios coupled with the restrictions for number of groups, number of dependent variables, and sample size ratio yield a total sample size ranging between 21 and 360. In addition, it leads to a ratio between total sample size and number of dependent variables, $\underline{N}$: $\underline{p}$, that varies between 7 and 60. Many current studies report this ratio (e.g., Coombs & Algina, in press; Algina & Tang, 1988).

Degree of heteroscedasticity ($\underline{d}$). A population in which the assumption of equal covariance matrices, homoscedasticity, holds is called uncontaminated. Such a population was simulated with a covariance matrix equal to a $\underline{p} \times \underline{p}$ identity matrix, I. A population in which the assumption of homoscedasticity is not met is called contaminated. The simulation of such violations was accomplished with a $\underline{p} \times \underline{p}$ diagonal population covariance matrix, **D**, with at least one diagonal element not equal to 1. In this study **D** was obtained by multiplying selected diagonal elements of the identity matrix by the square of the constant $\underline{d}$. The specific elements selected identified the dependent variables exhibiting heteroscedasticity (first, second, third, fourth, fifth, or sixth), while the value of $\underline{d}$ determined strength. The result was a diagonal matrix with $\underline{p}$ diagonals, some equal to 1 and the others equal to $\underline{d}$.

Shown in Table 3 are the covariance matrix forms used in the present study. Two levels of $\underline{d}$, $\underline{d} = 1.5$ and $\underline{d} = 3.0$ were used to simulate the degree of heteroscedasticity. These levels are similar to those used by Algina and Coombs (in press), $\underline{d} = \sqrt{2}$ and $\underline{d} = 3$. Olson (1974) employed $\underline{d} = 2.0$, 3.0, and 6.0. Algina and Tang (1988) chose $\underline{d} = 1.5$, 2.0, 2.5, and 3.0. Tang and Algina (1993) selected $\sqrt{1.5}$ and 3.0 for $\underline{d}$, while Algina and Oshima (1990) used 1.5 and 3.0. In this study, as in those cited, the smaller value simulates a low degree of heteroscedasticity, while the larger simulates a higher degree.

Table 3

Forms of Covariance Matrices

Matrix	$p = 3$	$p = 6$
D	$\text{Diag}(1, d^2, d^2)$	$\text{Diag}(1, 1, d^2, d^2, d^2, d^2)$
I	$\text{Diag}(1, 1, 1)$	$\text{Diag}(1, 1, 1, 1, 1, 1)$

Relationship between sample sizes and covariance matrices (s). This study incorporated investigations into both positive and negative relationships between sample sizes and covariance matrices (the positive and negative conditions). In the positive condition (denoted $s = 0$) the larger sample size was paired with **D**, the smaller with **I**. In the negative condition (denoted $s = 1$) the pairings were reversed, with the smaller samples paired with **D** and the larger with **I**. Shown in Table 4 is a summary of the relationship between sample sizes and covariance matrices.

Design Layout for Robustness Study

The sample sizes were determined by the levels of k, p, r, the sample size ratio F, and the relationship among the sample sizes. These sample sizes are summarized in Table 5 and Table 6. The range of sample sizes is broad enough to address theoretical issues, since it both overlaps with and extends the ranges of values used in previous studies addressing such issues. It is also braod enough to be useful to practitioners, since it was generated using values of variables (factors) that represent common ranges of usage. The 48 condition combinations in these tables were crossed with two distributions, two levels of heteroscedasticity, and two relationships between sample sizes and covariance matrices, resulting in 384

Table 4

Relationship between Sample Sizes and Covariance Matrices

$k = 3$

Sample Size Ratios			Relationship	
n_1 :	n_2 :	n_3	Positive	Negative
1	1	1.5	IID	DDI
1	1	2	IID	DDI
1	1.5	1.5	IDD	DII
1	2	2	IDD	DII

$k = 6$

Sample Size Ratios						Relationship	
n_1 :	n_2 :	n_3 :	n_4 :	n_5 :	n_6	Positive	Negative
1	1	1	1	1.5	1.5	IIIIDD	DDDDII
1	1	1	1	2	2	IIIIDD	DDDDII
1	1	1.5	1.5	1.5	1.5	IIDDDD	DDIIII
1	1	2	2	2	2	IIDDDD	DDIIII

experimental conditions upon which to base comparisons of Type I error rate for competing test criteria.

Simulation Procedure for the Robustness Study

The simulation was conducted as 384 separate runs, one for each combination of conditions described in the robustness study design layout, with exactly 20,000 replications per condition. For each condition, the performance of the Pillai-Bartlett V, Johansen J, Coombs-Algina U_1^*, Coombs-Algina U_2^*, Coombs- Algina L^*, and Coombs-Algina V^* tests were evaluated using the generated data.

Table 5

Sample Sizes for $\underline{k} = 3$

$\underline{p}$	$\underline{r}$	$\underline{n}_1$	$\underline{n}_2$	$\underline{n}_3$
3	2	6	6	9
		6	6	12
		6	9	9
		6	12	12
	4	12	12	18
		12	12	24
		12	18	18
		12	24	24
	6	18	18	27
		18	18	36
		18	27	27
		18	36	36
6	2	12	12	18
		12	12	24
		12	18	18
		12	24	24
	4	24	24	36
		24	24	48
		24	36	36
		24	48	48
	6	36	36	54
		36	36	72
		36	54	54
		36	72	72

For the $\underline{i}$th sample, an $\underline{n}_i \times \underline{p}$ ($\underline{i} = 1, 2, 3, \ldots \underline{k}$) matrix of uncorrelated pseudo-random observations was generated (using PROC IML in SAS) by selecting numbers from the target distribution, normal or exponential. When the target distribution was an

exponential, the random observations on each of the $\underline{p}$ variates were standardized using the population expected value and standard deviation. Hence, within each uncontaminated population, all the $\underline{p}$ variates were identically distributed with mean equal to zero, variance equal to one, and all covariances among the $\underline{p}$ variates equal to zero.

Table 6

Sample Sizes for $\underline{k} = 6$

$\underline{p}$	$\underline{r}$	$\underline{n}_1$	$\underline{n}_2$	$\underline{n}_3$	$\underline{n}_4$	$\underline{n}_5$	$\underline{n}_6$
3	2	6	6	6	6	9	9
		6	6	6	6	12	12
		6	6	9	9	9	9
		6	6	12	12	12	12
	4	12	12	12	12	18	18
		12	12	12	12	24	24
		12	12	18	18	18	18
		12	12	24	24	24	24
	6	18	18	18	18	27	27
		18	18	18	18	36	36
		18	18	27	27	27	27
		18	18	36	36	36	36
6	2	12	12	12	12	18	18
		12	12	12	12	24	24
		12	12	18	18	18	18
		12	12	24	24	24	24
	4	24	24	24	24	36	36
		24	24	24	24	48	48
		24	24	36	36	36	36
		24	24	48	48	48	48
	6	36	36	36	36	54	54
		36	36	36	36	72	72
		36	36	54	54	54	54
		36	36	72	72	72	72

Each $\underline{n}_i \times \underline{p}$ matrix of observations corresponding to a contaminated population was post multiplied by an appropriate **D** to simulate dispersion heteroscedasticity.

For each replication the data were analyzed using the Pillai-Bartlett $\underline{V}$, Johansen $\underline{J}$, Coombs-Algina $\underline{U}_1^*$, Coombs-Algina $\underline{U}_2^*$, Coombs-Algina $\underline{L}^*$, and Coombs-Algina $\underline{V}^*$ tests. The proportion of the 20,000 replications that yielded significant results at $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$ were recorded. These proportions served as estimates for the actual Type I error rates for the experimental runs.

Design Factors for Power Analysis

In addition to the seven factors incorporated into the design to investigate robustness of tests regarding Type I error rate, an additional factor was used to examine power. This factor was borrowed from Olson (1974) and is called noncentrality structure. Noncentrality structure is closely associated with the idea of noncentrality parameter.

The noncentrality parameter is a standardized measure of the differences present among population mean vectors. As such it is a measure of effect size. Schatzoff (1966) defined the noncentrality function as the trace of matrix **G**, tr(**G**), where

$$G = HV^{-1}.$$

V is the population covariance matrix and

$$H = \sum_{i=1}^{k} n_i(\mu_i - \mu)(\mu_i - \mu)' .$$

In the formula for H, μ_i is the population mean vector for the ith group, μ is the grand mean vector, and n_i is the sample size in the ith group. The matrix **G** is the multivariate extension of expressing mean difference in terms of standard deviations. Olson (1974) argues for substituting **I** for **V**, the population covariance matrix. Doing so allows for a test's ability to detect a given group-mean difference without assumption violations to be compared with its ability to detect the same difference when assumptions are violated.

In practice the noncentrality parameter tr(G) can be estimated by ($N - k$) times the Hotelling-Lawley trace. Olson (1974) used three levels in his study: 10, 40, and 90. However, in his analysis he cut these levels back to only one, 40. He observed that the problem of condensing results into a comprehensible package was complicated by the large number of factor combinations, and that it was urgent to cut back factors that were straightforward in their effect to one or two levels. Since an increase in the noncentrality parameter clearly increases power, comparisons between tests is best facilitated by using few levels of this factor. Olson cited a parameter of tr(G) = 40 as being well removed from the null case and yet not so high as to mask the effects. This study uses 40 as the value of the noncentrality parameter.

Noncentrality structure refers to the allocation of mean differences among various populations and among the various dependent variables. Two noncentrality structures were used in the present study, two of the three structures described by Olson (1974). One is a concentrated structure, whereas the second is a diffuse structure.

In the concentrated structure one group differs from the other $(\underline{k}-1)$ groups on a single variate. This structure was simulated by setting the mean vector of the first population equal to $(\underline{kc}, 0, 0)$ and all other mean vectors equal to (0, 0, 0) for $\underline{p} = 3$. For $\underline{p} = 6$ the first mean vector was $(\underline{kc}, 0, 0, 0, 0, 0)$, while all others were (0, 0, 0, 0, 0, 0). The value of $\underline{c}$ was determined by setting the noncentrality parameter tr(G) equal to the sum of the eigenvalues, $\underline{nk}(\underline{k}-1)\underline{c}^2$ (Olson, 1974).

The diffuse noncentrality structure is one in which each group differs from the others on a single dimension. This structure is simulated by setting all elements in each population mean vector equal to zero except the $\underline{i}$th element which is set equal to $\underline{kc}$ for all values of $\underline{i}$ from 1 to min($\underline{p}$, $\underline{k}$). For $\underline{k} = 6$ populations and $\underline{p} = 3$ dependent variables, the six population mean vectors are $(\underline{kc}, 0, 0)$, $(0, \underline{kc}, 0)$, $(0, 0, \underline{kc})$, (0, 0, 0), (0, 0, 0), and (0, 0, 0). For $\underline{k} = 3$ groups and $\underline{p} = 6$ dependent variables, the population mean vectors are $(\delta, 0, 0, 0, 0, 0)$, $(0, \delta, 0, 0, 0, 0)$, and $(0, 0, \delta, 0, 0, 0)$. The value of $\underline{c}$ was found by setting the noncentrality parameter equal to $(\underline{p}-1)\underline{nk}^2\underline{c}^2 + \underline{nk}(\underline{k}-\underline{p})\underline{c}^2$ when $\underline{k} > \underline{p}$ and $(\underline{k}-1)\underline{nk}^2\underline{c}^2$ when $\underline{k} \leq \underline{p}$ (Olson, 1974).

Design Layout for Power Analysis

Each of the 384 experimental conditions generated for the comparisons of Type I error rate was crossed with two noncentrality structures for the purpose of evaluating power, yielding 768 experimental conditions upon which to compare the competing statistical tests.

Simulation Procedure for the Power Analysis

A simulation was conducted as 768 separate runs, one for each of the condition combinations described in the power analysis design layout, with 20,000 replications per condition combination. Distributions and heteroscedasticity were simulated using PROC IML in SAS as in the robustness study.

For each replication, the data were analyzed using the Pillai-Bartlett $\underline{V}$, Johansen $\underline{J}$, Coombs-Algina $\underline{U}_1^*$, Coombs-Algina $\underline{U}_2^*$, Coombs-Algina $\underline{L}^*$, and Coombs-Algina $\underline{V}^*$ tests. The proportion of the 20,000 replications yielding significant results at $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$ were recorded. These proportions served as estimates of the power of the test for the various condition combinations under the specified noncentrality parameter and structure.

Summary

Two distribution types ($\underline{DT}$ = normal or exponential), two levels of populations sampled ($\underline{k} = 3$ or 6), two levels of dependent variables ($\underline{p} = 3$ or 6), three levels of the ratio between smallest sample size and number of dependent variables ($\underline{r} = 2$, 4, or 6), four levels of the sample size ratio, two levels of degree of heteroscedasticity ($\underline{d} = 1.5$ or 3.0), and two levels of the relationship between sample sizes and covariance matrices (positive and negative) combine to give 384 experimental conditions upon which to base conclusions regarding control of Type

I error rate. For each condition, one noncentrality parameter (40) and two noncentrality structures (concentrated and diffuse) combine to produce 768 conditions in which to compare the powers of tests. The Pillai-Bartlett $\underline{V}$, Johansen $\underline{J}$, Coombs-Algina $\underline{U}_1^*$, Coombs-Algina $\underline{U}_2^*$, Coombs-Algina $\underline{L}^*$, and Coombs-Algina $\underline{V}^*$ tests were applied to each of these experimental conditions. Generalizations of the behavior of these tests will be based upon the collective results of the experimental conditions.

Chapter 4
Results

In this chapter estimated Type I error rates and power levels under various condition combinations are presented and discussed for tests performed at the .05 level of significance.

Type I Error Rate Results

Figures 1-6 depict the distributions of the estimated Type I error rates for the Pillai-Bartlett, Johansen, Coombs-Algina $\underline{U}_1^*$, Coombs-Algina $\underline{U}_1^*$, Coombs-Algina $\underline{L}^*$, and Coombs-Algina $\underline{V}^*$ tests. Table 7 further describes these distributions by reporting five percentiles for each of the tests. Five percentils are provided for each test statitic. The third entry in the first row, for example, reports that 50% (percentile = 50) of the estimated Type I error rates equal .0546 or less. In terms of controlling Type I error rates, these results indicate (a) the Pillai-Bartlett and Johansen tests are similar in performance with the Pillai-

Table 7

Percentiles for Estimated Type I Error Rate

Test Criterion	Percentile				
	0	25	50	75	100
Pillai-Bartlett	.0092	.0306	.0546	.1052	.3455
Johansen	.0457	.0601	.0816	.1235	.4469
Coombs-Algina $\underline{U}_1^*$	.0287	.0496	.0544	.0614	.1004
Coombs-Algina $\underline{U}_2^*$	.0222	.0451	.0511	.0578	.0940
Coombs-Algina $\underline{L}^*$	.0193	.0440	.0504	.0558	.0904
Coombs-Algina $\underline{V}^*$	.0084	.0363	.0458	.0512	.0802

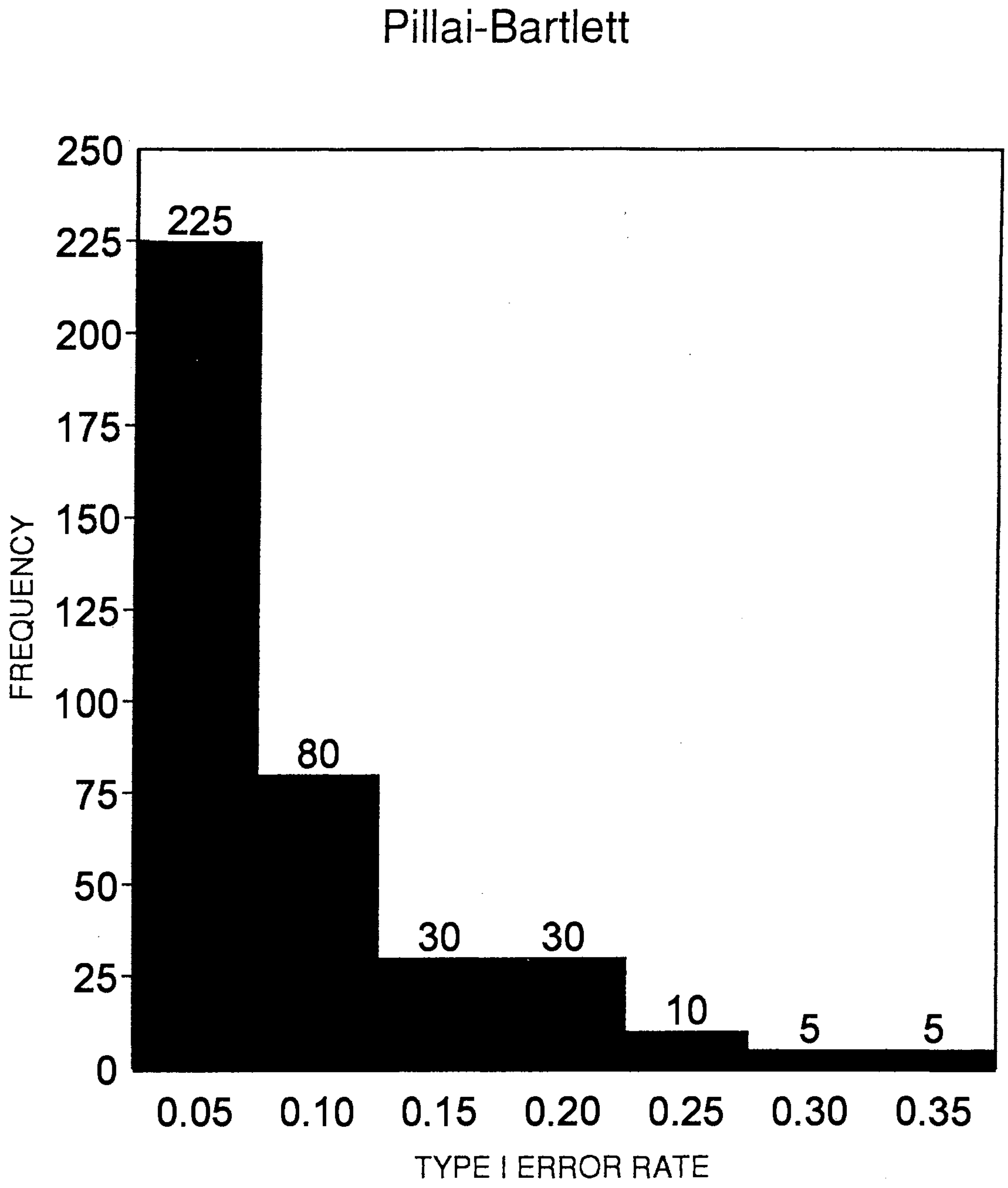


Figure 1. Frequency histogram of estimated Type I error rates for the Pillai-Bartlett test.

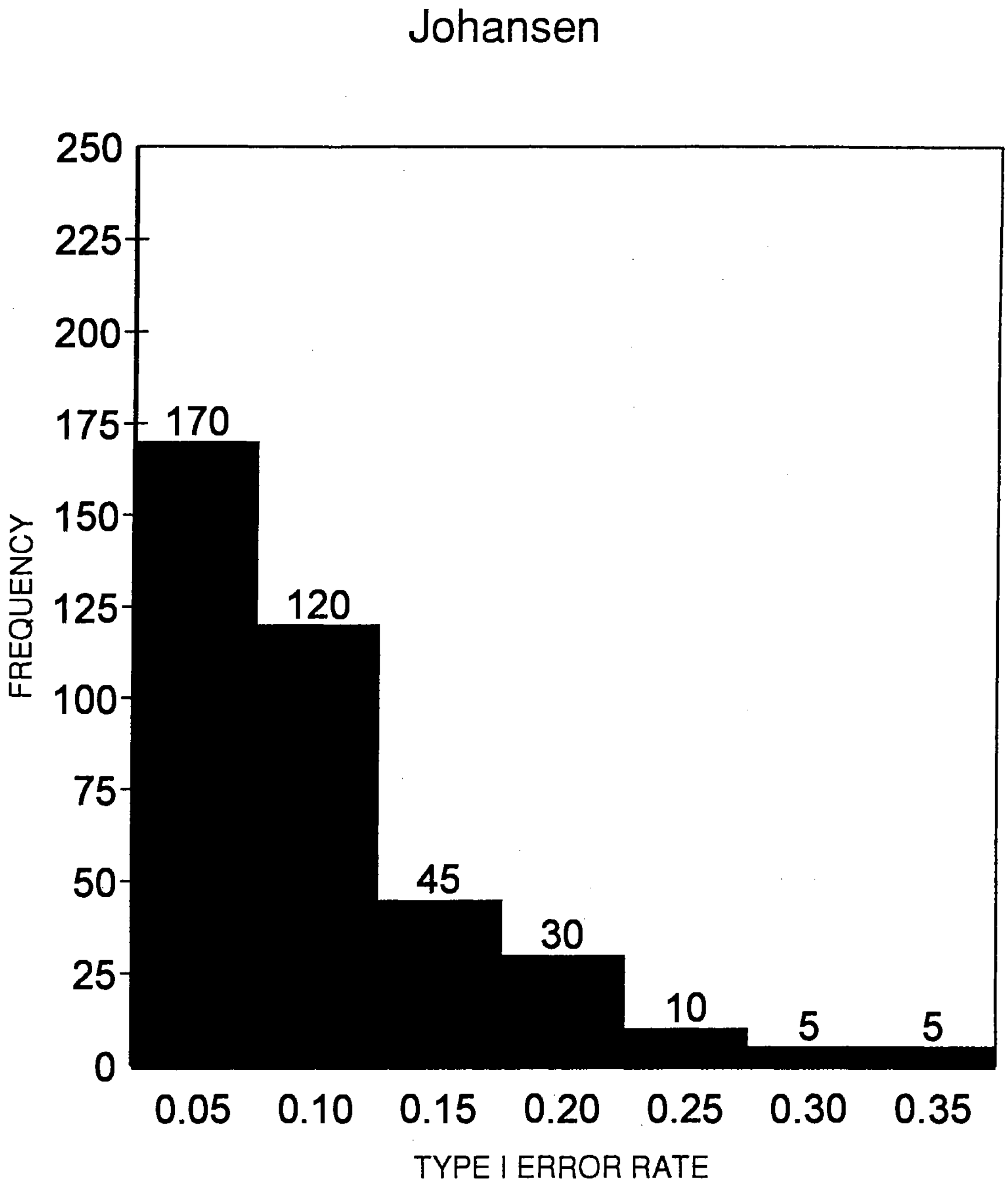


Figure 2. Frequency histogram of estimated Type I error rates for the Johansen test.

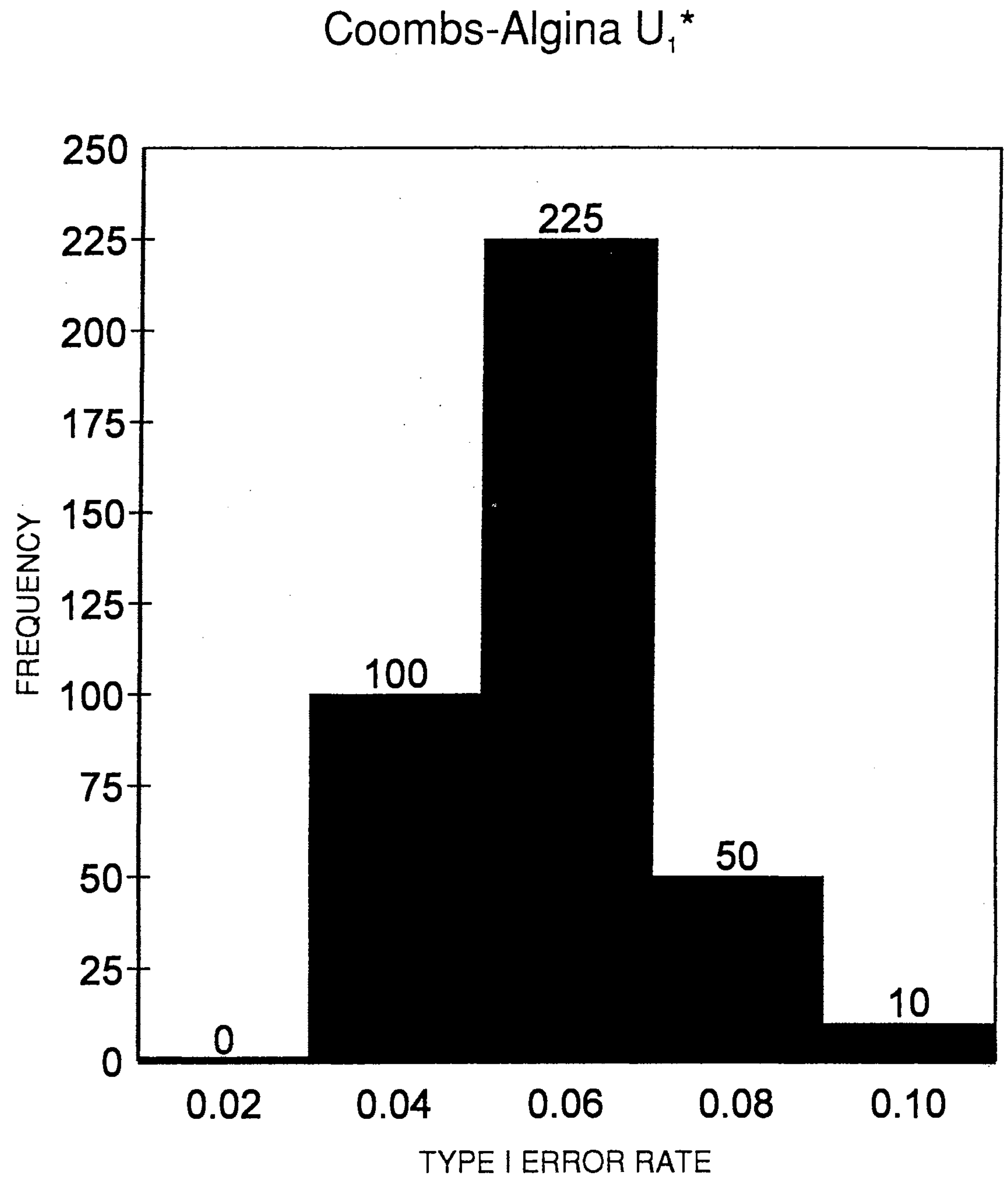


Figure 3. Frequency histogram of estimated Type I error rates for the Coombs-Algina U_1^* test.

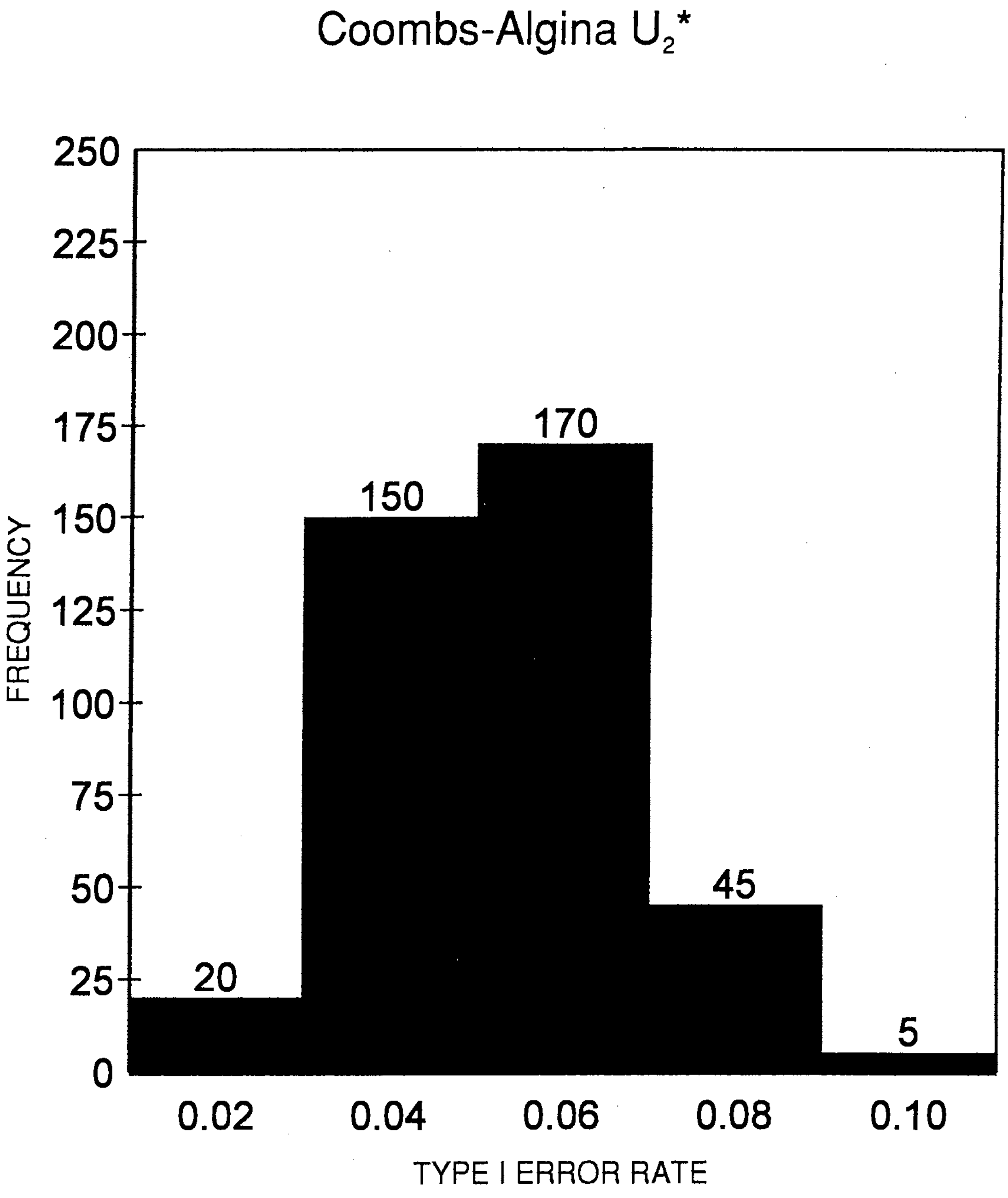


Figure 4. Frequency histogram of estimated Type I error rates for the Coombs-Algina U_2* test.

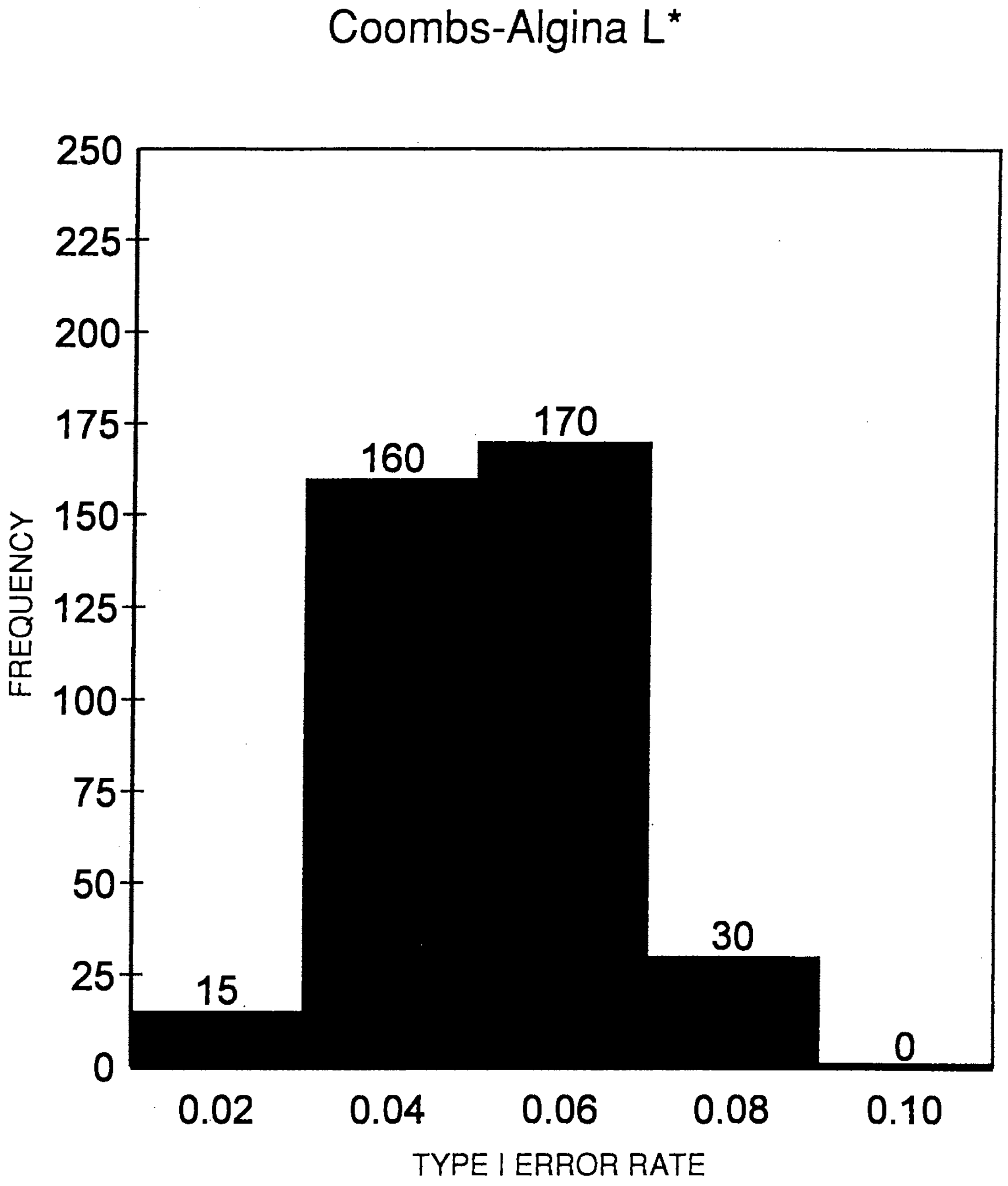


Figure 5. Frequency histogram of estimated Type I error rates for the Coombs-Algina L* test.

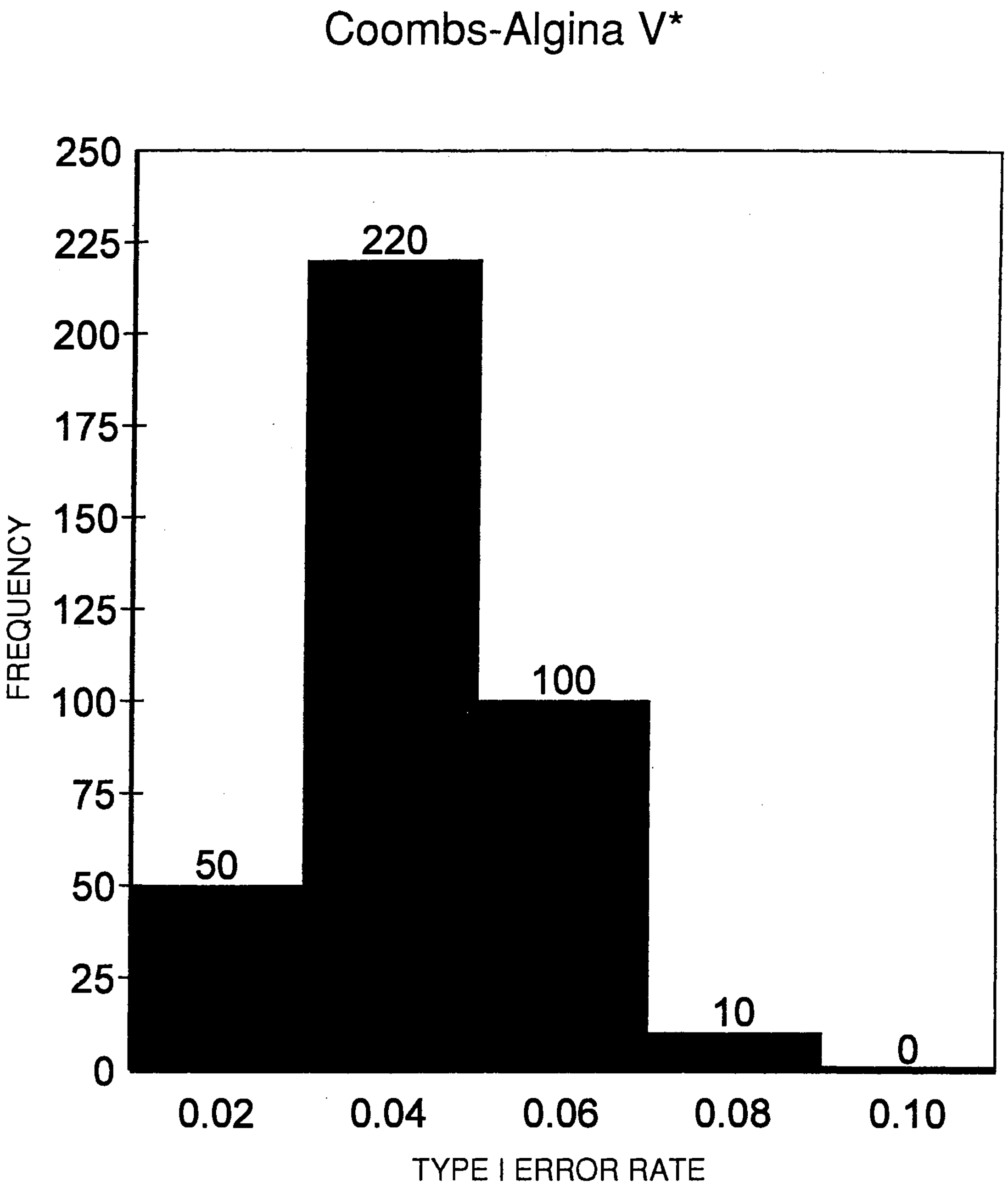


Figure 6. Frequency histogram of estimated Type I error rates for the Coombs-Algina V* test.

Bartlett's having somewhat lower levels, (b) the four Coombs-Algina tests are similar, and (c) the performance of each of the four Coombs-Algina tests is superior to that of either the Pillai-Bartlett test or the Johansen test. Examination of the histograms indicates the order of the estimated Type I error rates for the four Coombs-Algina tests from smallest to largest to be: $\underline{V}^*$, $\underline{L}^*$, $\underline{U}_2^*$, $\underline{U}_1^*$. These differences, however, appear to be slight based on examination of either the histograms or percentiles. This is a result that was expected, given that the four statistics are all functions of the same matrix, $\mathbf{HM}^{-1}$.

The Pillai-Bartlett test satisfies Bradley's (1978) liberal criterion ($.5\alpha \leq \hat{\tau} \leq 1.5\alpha$) in only about 58% of the conditions studied. Only about 44% of the estimated Type I error rates for the Johansen test fall in Bradley's interval. The failure of these tests to achieve nominal levels in such a large percentage of cases alone justifies eliminating them from consideration under the conditions considered in this study. On the other hand, all four Coombs-Algina tests have estimated Type I error rates that fall in Bradley's interval in nearly 90% of all cases. More specifically, the success rates are approximately 88% for $\underline{U}_1^*$, 86% for $\underline{U}_2^*$, 91% for $\underline{L}^*$, and 88% for $\underline{V}^*$.

Type I error rate was further analyzed using a split-plot analysis of variance model. Seven between factors (distribution type, number of groups sampled, number of dependent variables, sample size ratio form, ratio of smallest sample size to number of dependent variables, degree of heteroscedasticity, and the relationship between sample sizes and covariance matrices) and one within factor (test criterion) were included in the model. All main effects and two-way through seven-way interactions were considered. Because no replications were made for each combination of conditions in generating the data, the error term for both the between and within analyses was the mean squared error for the highest order interaction for that analysis. That is, the error term for the between analysis was the mean squared error for the effect $\underline{DT} \times \underline{k} \times \underline{p} \times \underline{F} \times \underline{r} \times \underline{d} \times \underline{s}$. For the within analysis the error term was the mean squared error for the interaction of those same factors with test criterion ($\underline{T}$).

Because the extremely large sample sizes resulted in a large number of statistically significant effects, practical significance was estimated using omega-squared ($\hat{\omega}^2$), which reports the proportion of total variance accounted for by an effect.

$$\hat{\omega}^2 = \frac{\hat{\delta}^2_{effect}}{\Sigma\hat{\delta}^2_{effect} + MS_{between} + MS_{within}}$$

where

$$\hat{\delta}^2_{effect} = \frac{df_{effect}(MS_{effect} - MS_{between})}{2304}\ .$$

The constant 2304 is the total number of conditions considered in the study (that is, the product of the levels of all factors examined).

Nineteen effects accounted for 90.3190% of the total variance. Fourteen of these were within factors that accounted for 78.991% and five were between factors that accounted for the remaining 11.328%. These effects and their $\hat{\omega}^2$ values appear in Table 8 and include all effects that accounted for at least 1% of the total variance. (More precisely, all $\hat{\omega}^2$ values that rounded to at least .009 were included.) Only one variable, $\underline{p}$ = number of dependent variables, failed to appear in at least one statistically and practically significant effect. All others were involved in one of the four effects that subsumed all others:

$$T \times F \times s \times d$$

$$T \times s \times r$$

$$T \times k \times r$$

$$T \times k \times DT.$$

The four-way interaction T × F × s × d and the ten significant (statistically and practically) effects it subsumes account for 62.938% of the total variance. The effect T × r × s and the four significant effects it subsumes account for 58.212% of the total variance. The effect T × k × r and the three effects it subsumes account for 29.657% of the total variance. And the effect T × k × DT and the four significant effects it subsumes account for 28.887% of the total

Table 8

Proportion of Variance in Estimated Type I Error Rate Accounted for by Statistically and Practically Significant Effects

	Effect	Proportion of Total Variance
Within	T × r × s	.00912
	T × F × d	.00923
	T × DT × k	.00924
	T × F × s × d	.01099
	T × F	.01991
	T × F × s	.02217
	T × d	.02568
	T × k × r	.02642
	T × k	.03853
	T × DT	.03926
	T × s × d	.04898
	T × r	.12506
	T × s	.19113
	T	.21419
		.78991
Between	DT	.008745
	F × s	.009939
	k	.017433
	d	.034541
	s	.042622
		.113280

variance. The sum of these percentages exceeds the 90.3190% accounted for by the statistically and practically significant effects since some effects such as test criterion (T) are subsumed more than once. These percentages are organized in Table 9.

Table 9

Proportion of Variance in Estimated Type I Error Rate Accounted for by Various Effects by Group

Group	Effect	Proportion of Total Variance
T × F × s × d	T × F × s × d	.01099
	T × F × d	.00923
	T × F × s	.02217
	T × s × d	.04898
	F × s	.00994
	T × F	.01991
	T × d	.02568
	T × s	.19113
	d	.03454
	s	.04262
	T	.21419
		.62938
T × r × s	T × r × s	.00912
	T × r	.12506
	T × s	.19113
	s	.04262
	T	.21419
		.58212
T × k × r	T × k × r	.02642
	T × k	.03853
	k	.01743
	T	.21419
		.29657
T × k × DT	T × k × DT	.00924
	T × DT	.03926
	DT	.00875
	k	.01743
	T	.21419
		.28887

Because test criterion was included in all four effects that subsumed all statistically and practically significant effects, the effects of the other variables and their interactions must be considered test by test.

Effects of Sample Size Ratio Form, Relationship Between Sample Sizes and Covariance Matrices, and Degree of Heteroscedasticity ($\underline{F}$, $\underline{s}$, and $\underline{d}$). The $\underline{T} \times \underline{F} \times \underline{s} \times \underline{d}$ interaction was examined by constructing all mean plots involving all combinations of sample size ratio form ($\underline{F}$), relationship between sample sizes and covariance matrices ($\underline{s}$), and degree of heteroscedasticity ($\underline{d}$) for each test. The plots appear in Figures 7-12. Figure 13 includes them all for ease of comparison.

The Johansen test is on the average always liberal with mean values of $\hat{\tau}$ ranging from .0791 ($\underline{s} = 0$, $\underline{d} = 1.5$, $\underline{F} = 5$) to .1567 ($\underline{s} = 1$, $\underline{d} = 3$, $\underline{F} = 3$). The Pillai-Bartlett test is liberal in the negative condition ($\underline{s} = 1$) and conservative in the positive condition ($\underline{s} = 0$) for all combinations of sample size ratio form and degree of heteroscedasticity. For both the Pillai-Bartlett and Johansen tests, estimated mean Type I error rate is larger in the negative condition ($\underline{s} = 1$) than in the positive condition ($\underline{s} = 0$). The result is reversed for the Coombs-Algina $\underline{V}^*$ test with mean $\hat{\tau}$s being larger is the positive condition than in the negative condition. For the other three Coombs-Algina tests the plots of the positive and negative conditions are disordinal indicating that the relative sizes of mean $\hat{\tau}$ in the positive and negative conditions change with the degree of heteroscedasticity and with sample size ratio form. All means for all Coombs-Algina tests fall within or nearly within Bradley's (1978) liberal criterion interval, which for the .05 level of significance is between .025 and .075.

For the Coombs-Algina $\underline{U}_1^*$ and $\underline{U}_2^*$ tests the patterns are the same. Estimated mean Type I error rate is larger in the positive condition for forms 2 and 3 (1:1:1.5, 1:1:1:1:1.5:1.5, 1:1:2, or 1:1:1:1:2:2). The tests are slightly liberal

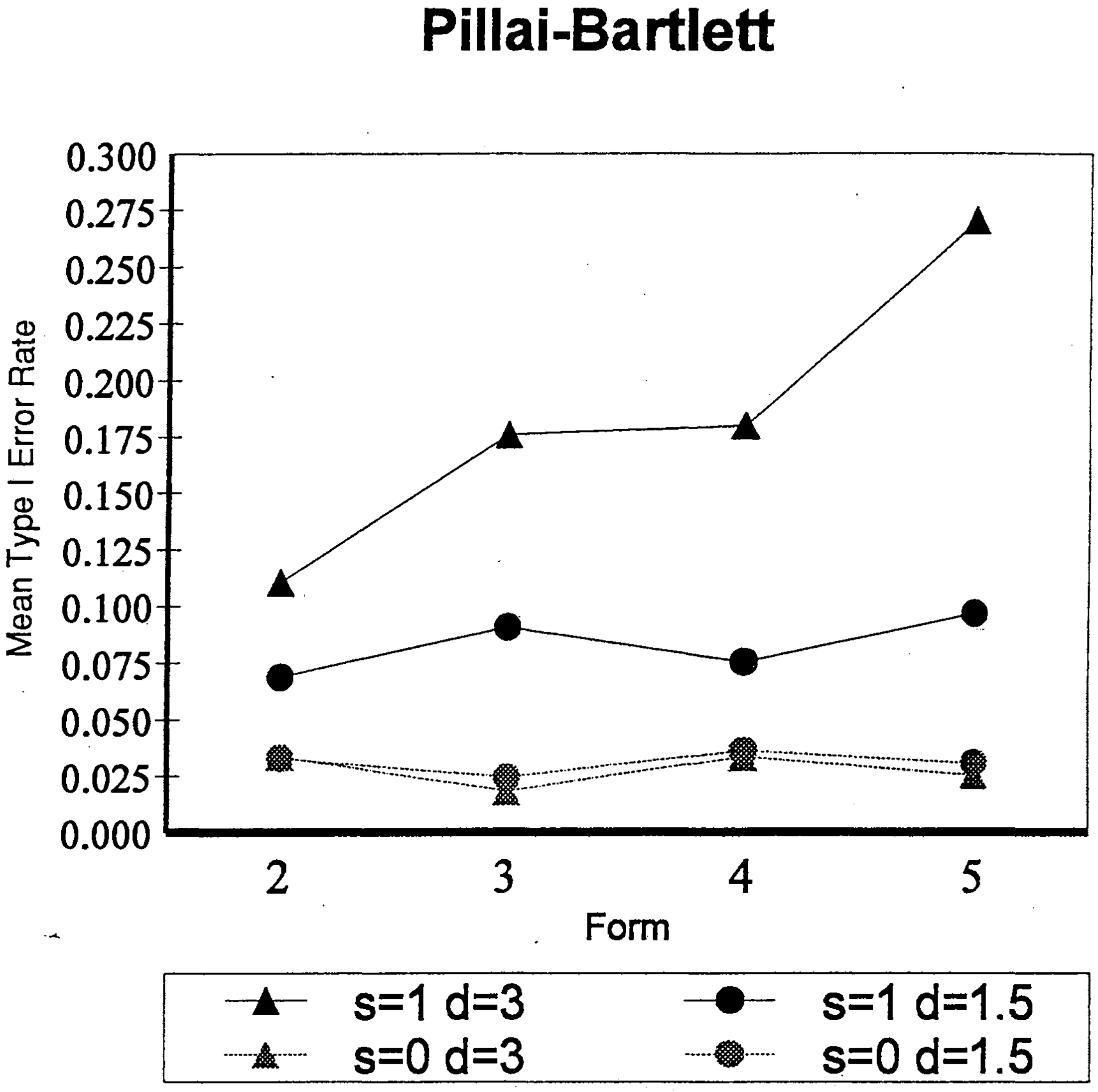


Figure 7. Estimated Type I error rates for combinations of sample size ratio form (F), relationship between sample size and covariance matrices (s) and degree of heteroscedasticity (d) for the Pillai-Bartlett test.

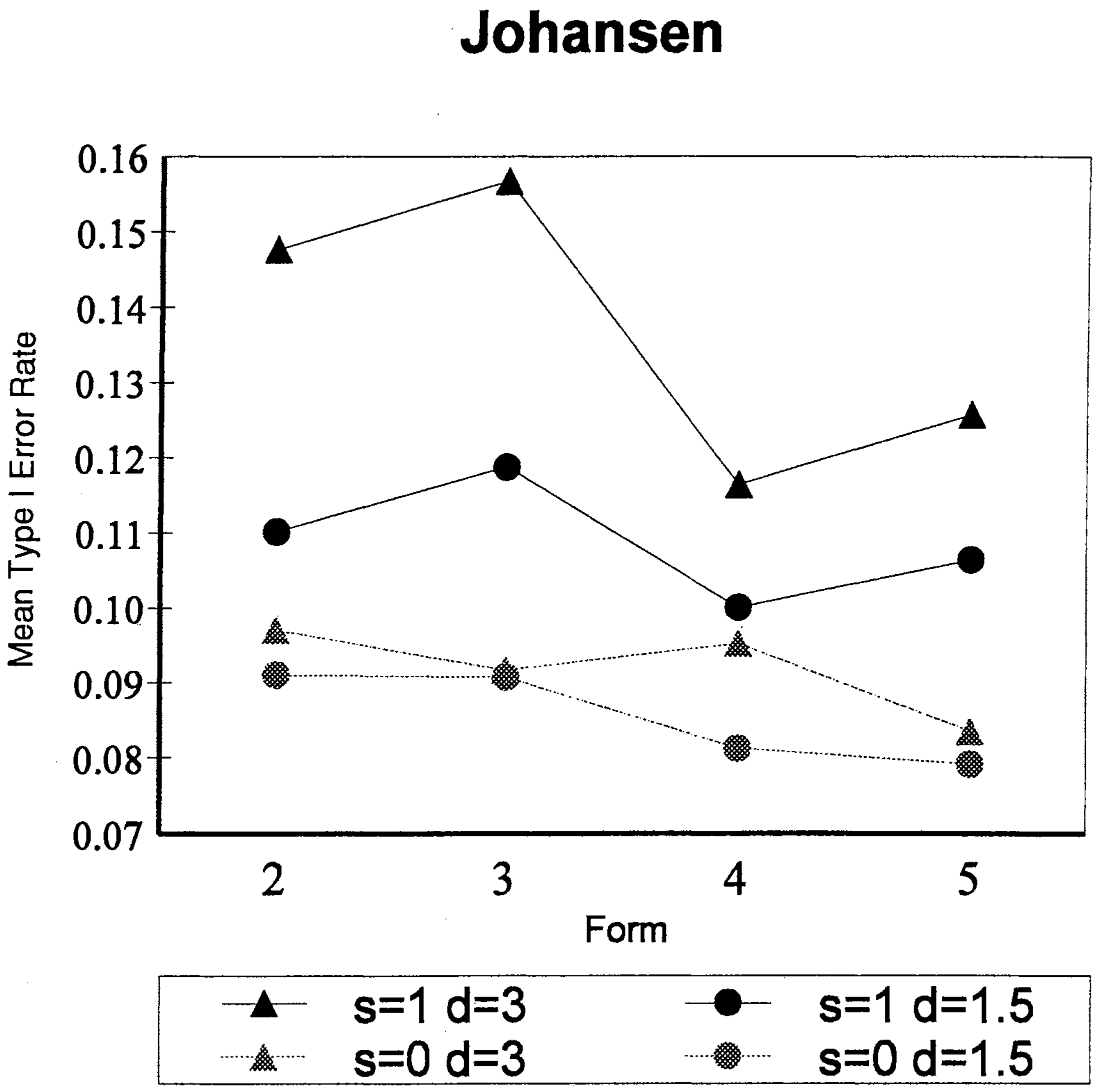


Figure 8. Estimated Type I error rates for combinations of sample size ratio form (F), relationship between sample size and covariance matrices (s) and degree of heteroscedasticity (d) for the Johansen test.

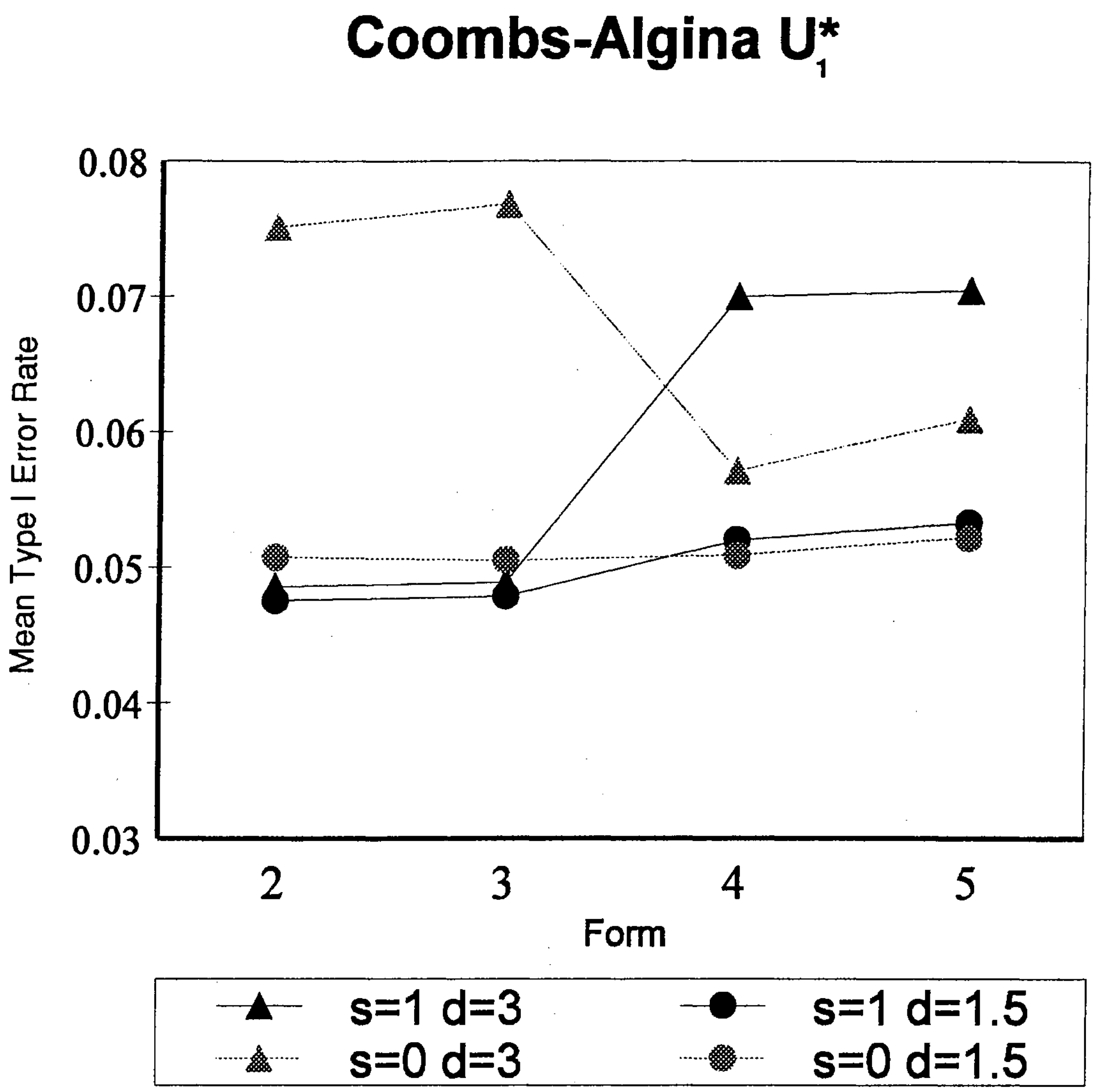


Figure 9. Estimated Type I error rates for combinations of sample size ratio form (F), relationship between sample size and covariance matrices (s) and degree of heteroscedasticity (d) for the Coombs-Algina U_1* test.

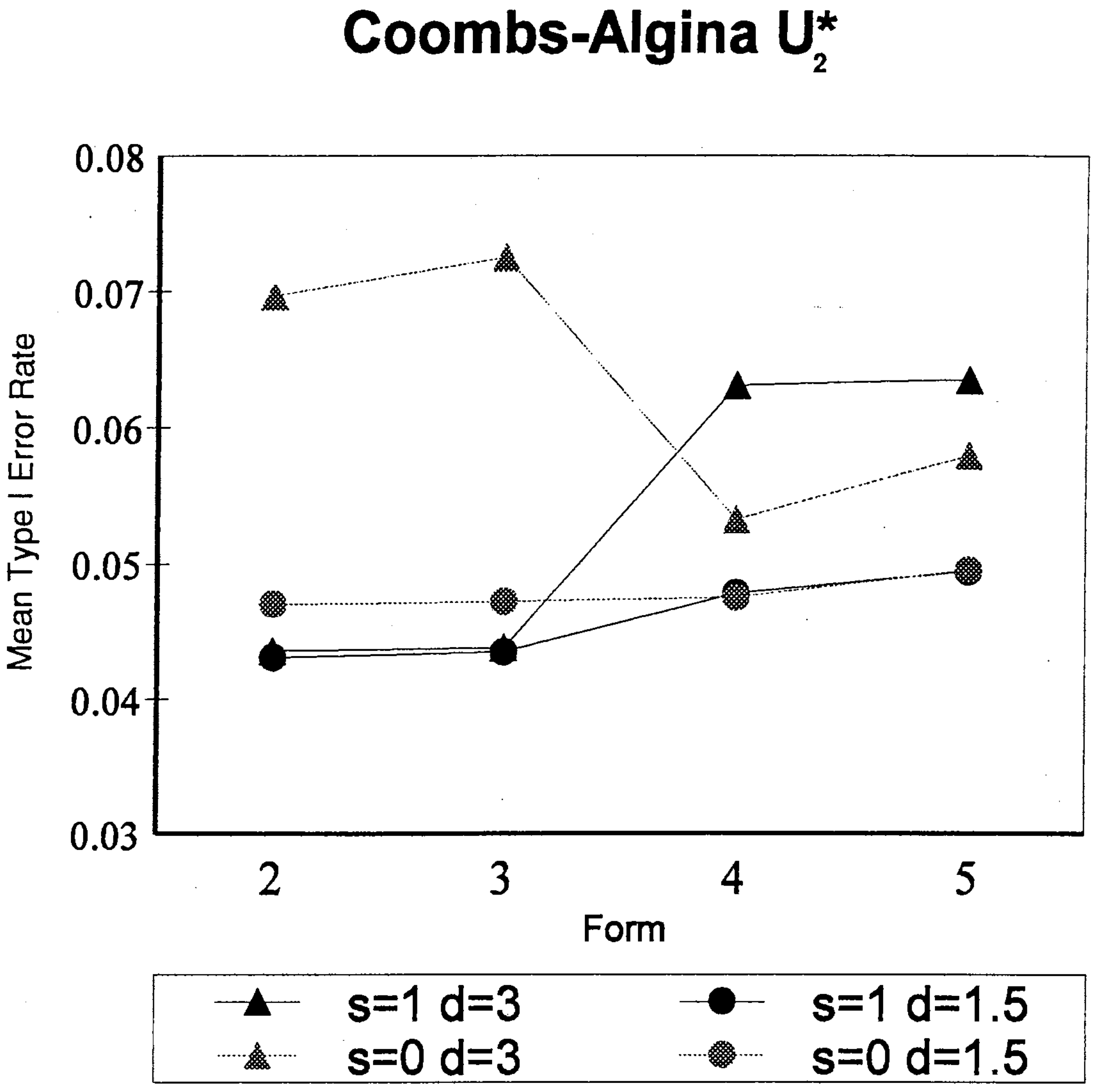


Figure 10. Estimated Type I error rates for combinations of sample size ratio form (F), relationship between sample size and covariance matrices (s) and degree of heteroscedasticity (d) for the Coombs-Algina U_2* test.

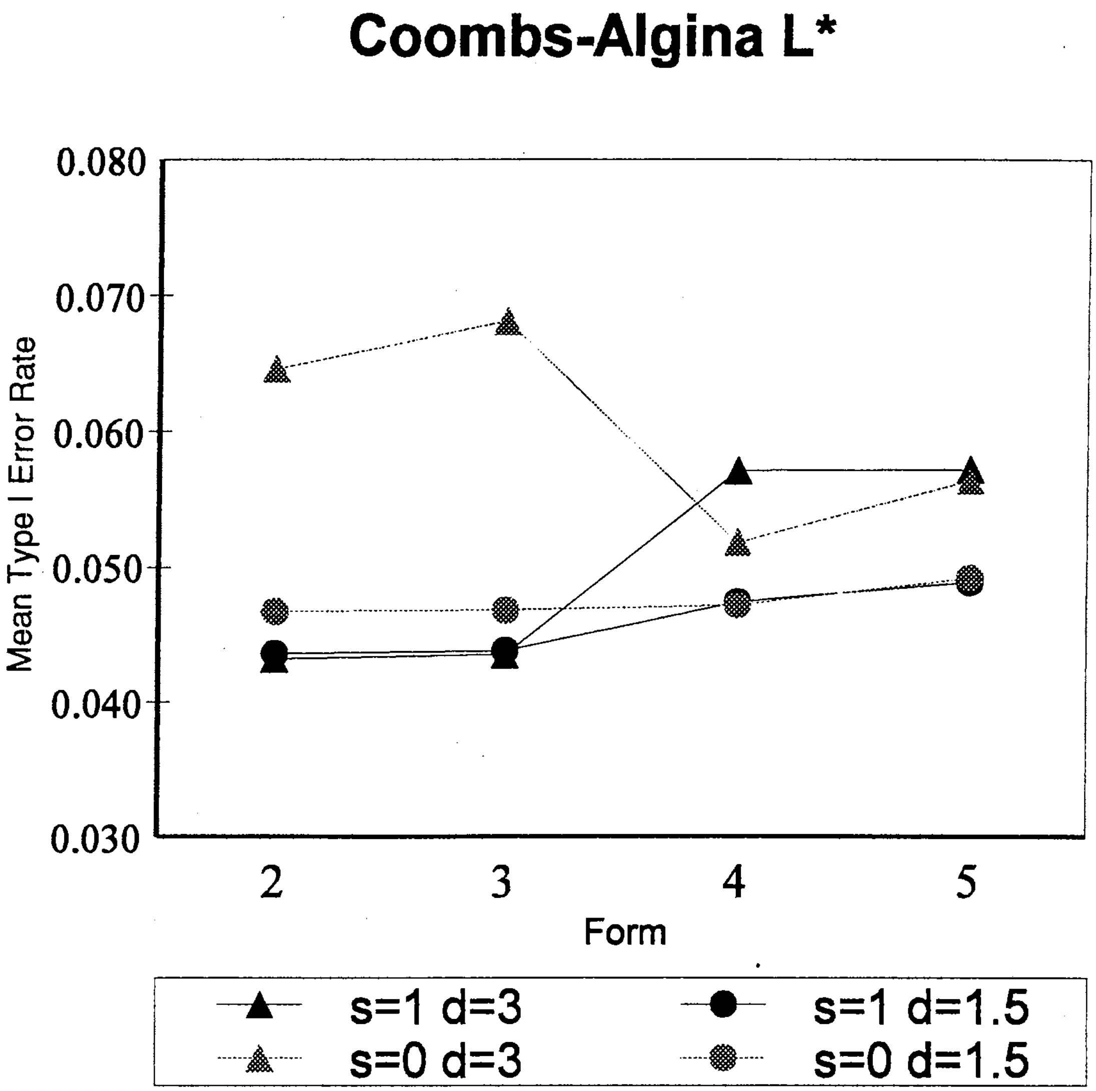


Figure 11. Estimated Type I error rates for combinations of sample size ratio form (F), relationship between sample size and covariance matrices (s) and degree of heteroscedasticity (d) for the Coombs-Algina L* test.

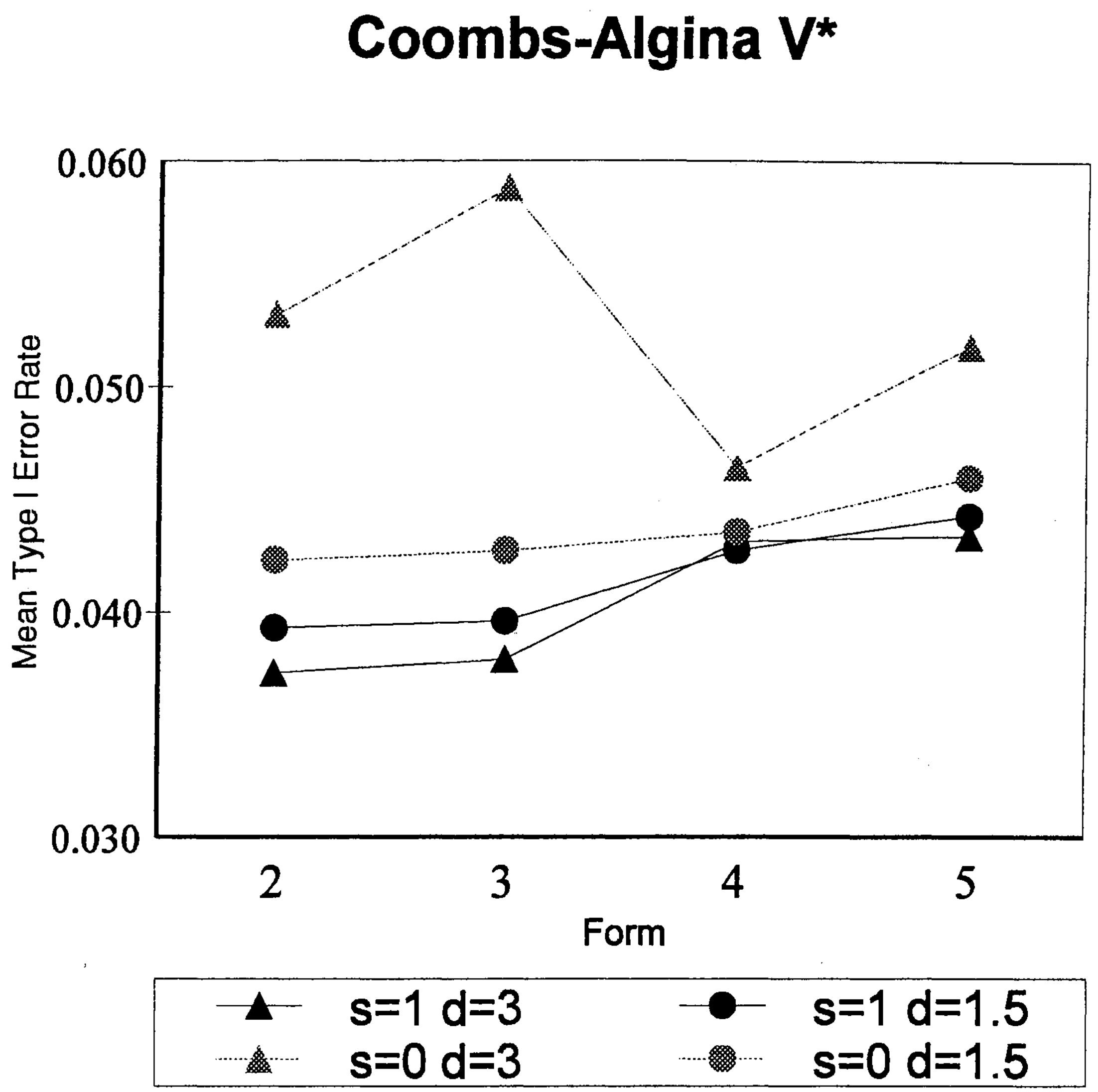
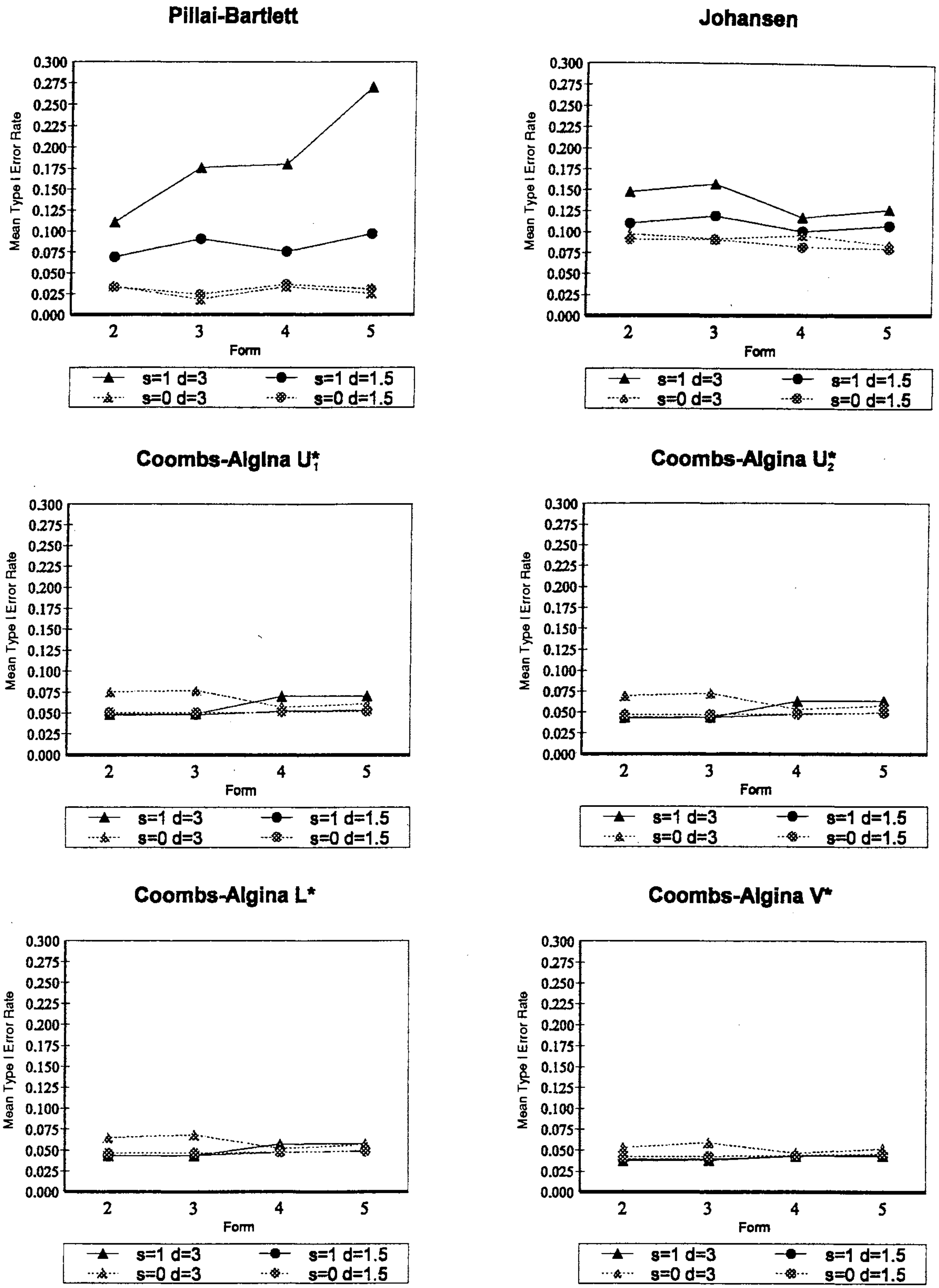


Figure 12. Estimated Type I error rates for combinations of sample size ratio form (F), relationship between sample size and covariance matrices (s) and degree of heteroscedasticity (d) for the Coombs-Algina V* test.



Pillai-Bartlett

Johansen

Coombs-Algina U_1^*

Coombs-Algina U_2^*

Coombs-Algina L*

Coombs-Algina V*

Figure 13. Estimated Type I error rates for combinations of sample size ratio form (F), relationship between sample size and covariance matrices (s) and degree of heteroscedasticity(d) for six tests.

when $\underline{d} = 3$ and $\underline{s} = 0$ (positive condition), but are near nominal in all other cases of forms 2 and 3. For forms 4 and 5 (1:1.5:1.5, 1:1:1.5:1.5:1.5:1.5, 1:2:2, or 1:1:2:2:2:2) $\underline{U}_1^*$ and $\underline{U}_2^*$ tend to yield higher estimated mean Type I error rates for the negative condition for both $\underline{d} = 3$ and $\underline{d} = 1.5$. The rates are higher when $\underline{d} = 3$ than when $\underline{d} = 1.5$ for forms 4 and 5.

Mean Type I error rate estimates obtained in the Coombs-Algina $\underline{L}^*$ test are larger in the positive condition for forms 2 and 3. For those forms mean $\hat{\tau}$ is in the upper range of Bradley's (1978) interval when $\underline{d} = 3$ and $\underline{s} = 0$ (positive condition) and near nominal in all other conditions. For forms 4 and 5 estimated mean Type I error rate for the Coombs-Algina $\underline{L}^*$ test is higher when $\underline{d} = 3$ than when $\underline{d} = 1.5$.

For all six tests $\underline{d} = 1.5$ yields a much flatter plot than does $\underline{d} = 3.0$, indicating that Type I error rate varies more as the degree of covariance matrix inequality increases. This tendency is especially true for the Coombs-Algina tests.

The Coombs-Algina tests show smaller differences in estimated mean Type I error rates between the positive and negative conditions for forms 4 and 5 than do either the Pillai-Bartlett test or Johansen test.

Effects of Relationship Between Sample Sizes and Covariance Matrices and Ratio Between Smallest Sample Size and Number of Dependent Variables ($\underline{s}$ and $\underline{r}$). Cell mean plots of combinations of the relationship between sample sizes and covariance matrices ($\underline{s}$) and ratio of smallest sample size to the number of dependent variables ($\underline{r}$) appear separately in Figures 14-19 and as a group in Figure 20. They reveal the Pillai-Bartlett test to be liberal in the negative condition ($\underline{s} = 1$) and near the low end of Bradley's (1978) interval in the positive condition ($\underline{s} = 0$). The value of $\underline{r}$ has little, if any, effect on Type I error rate for the Pillai-Bartlett test. Under the conditions considered the Johansen test is

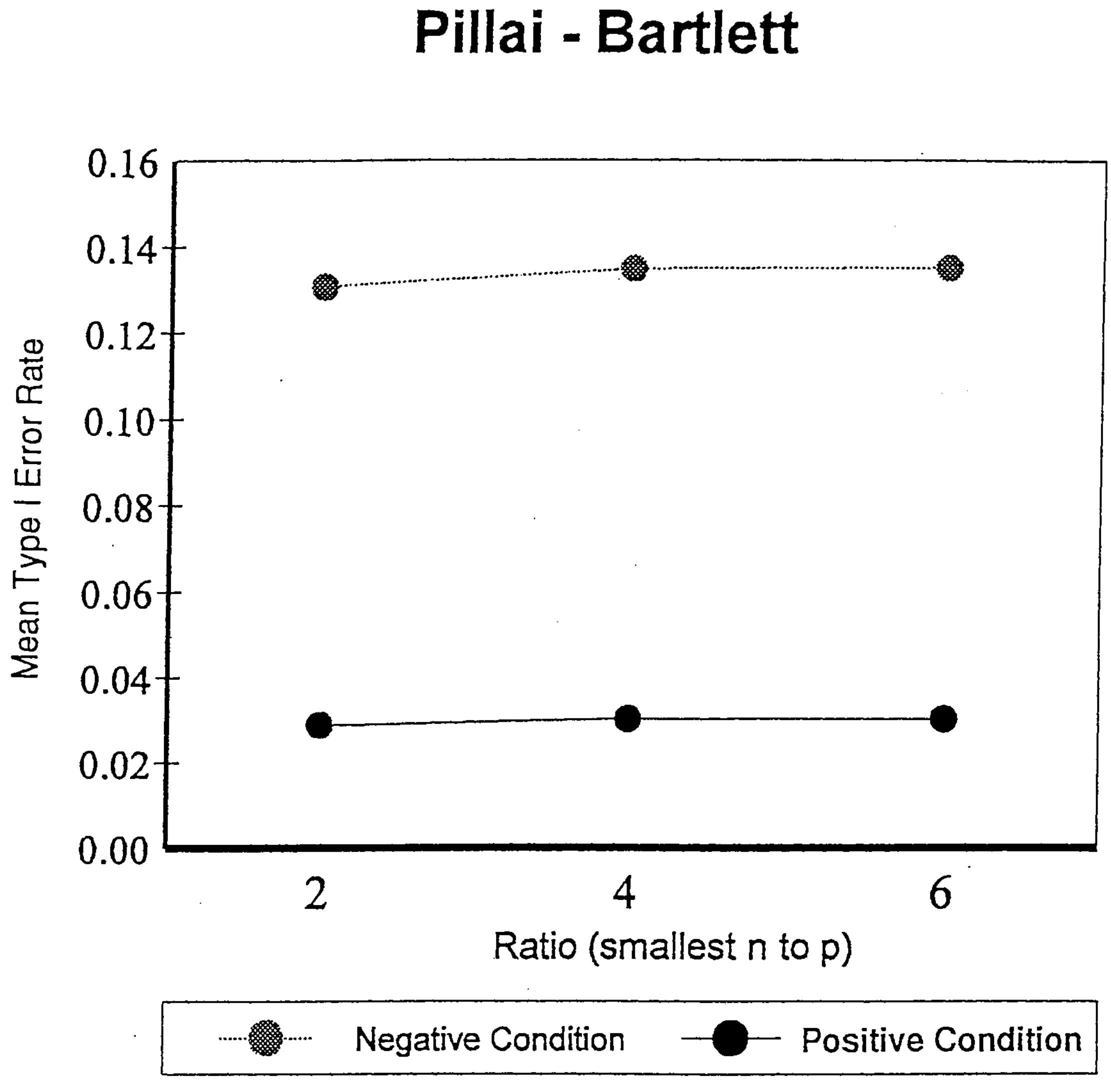


Figure 14. Estimated Type I error rates for combinations of ratios of smallest sample size to number of dependent variables (r) and relationship between sample size and covariance matrices (s) for the Pillai-Bartlett test.

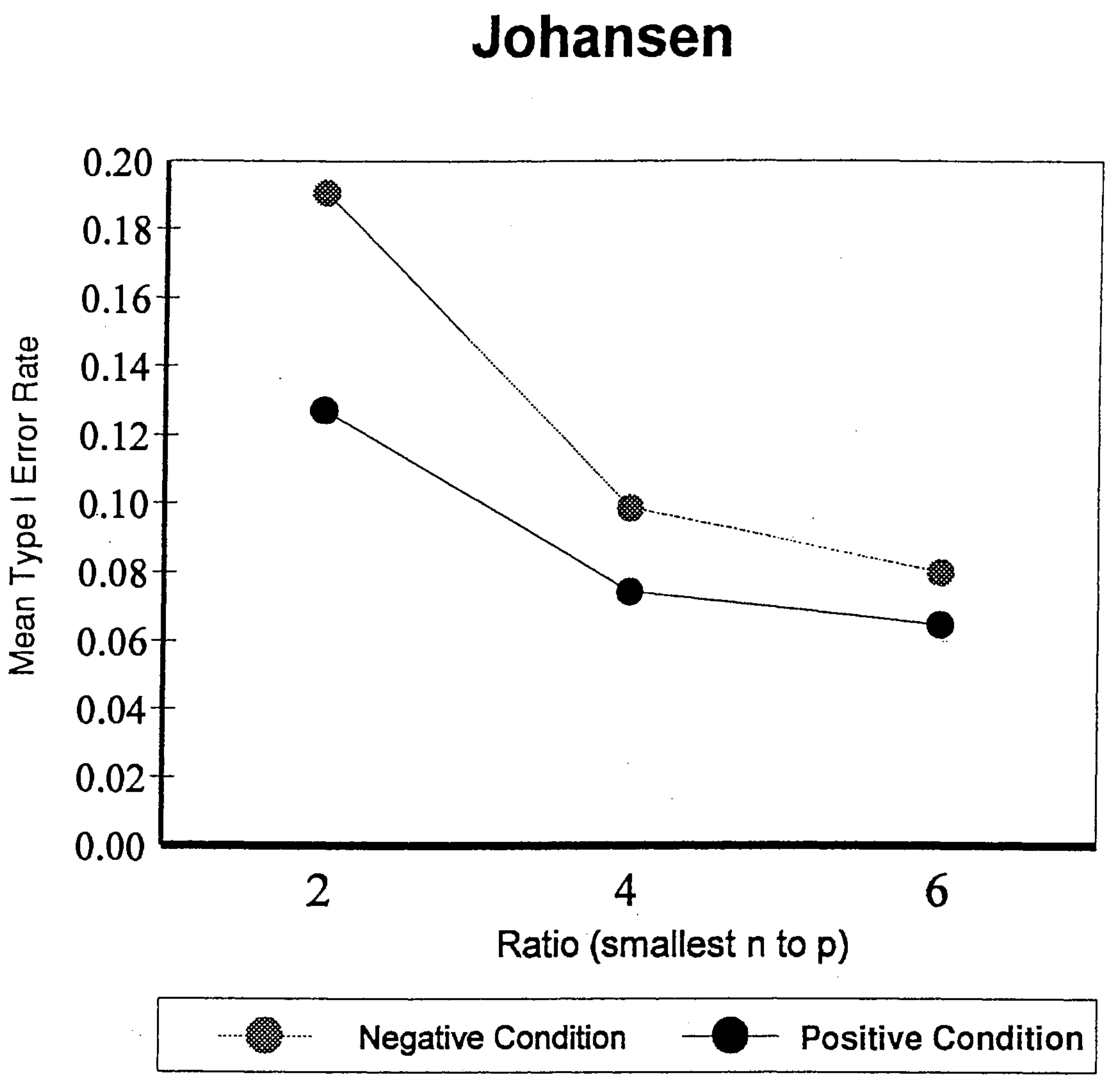


Figure 15. Estimated Type I error rates for combinations of ratios of smallest sample size to number of dependent variables (r) and relationship between sample size and covariance matrices (s) for the Johansen test.

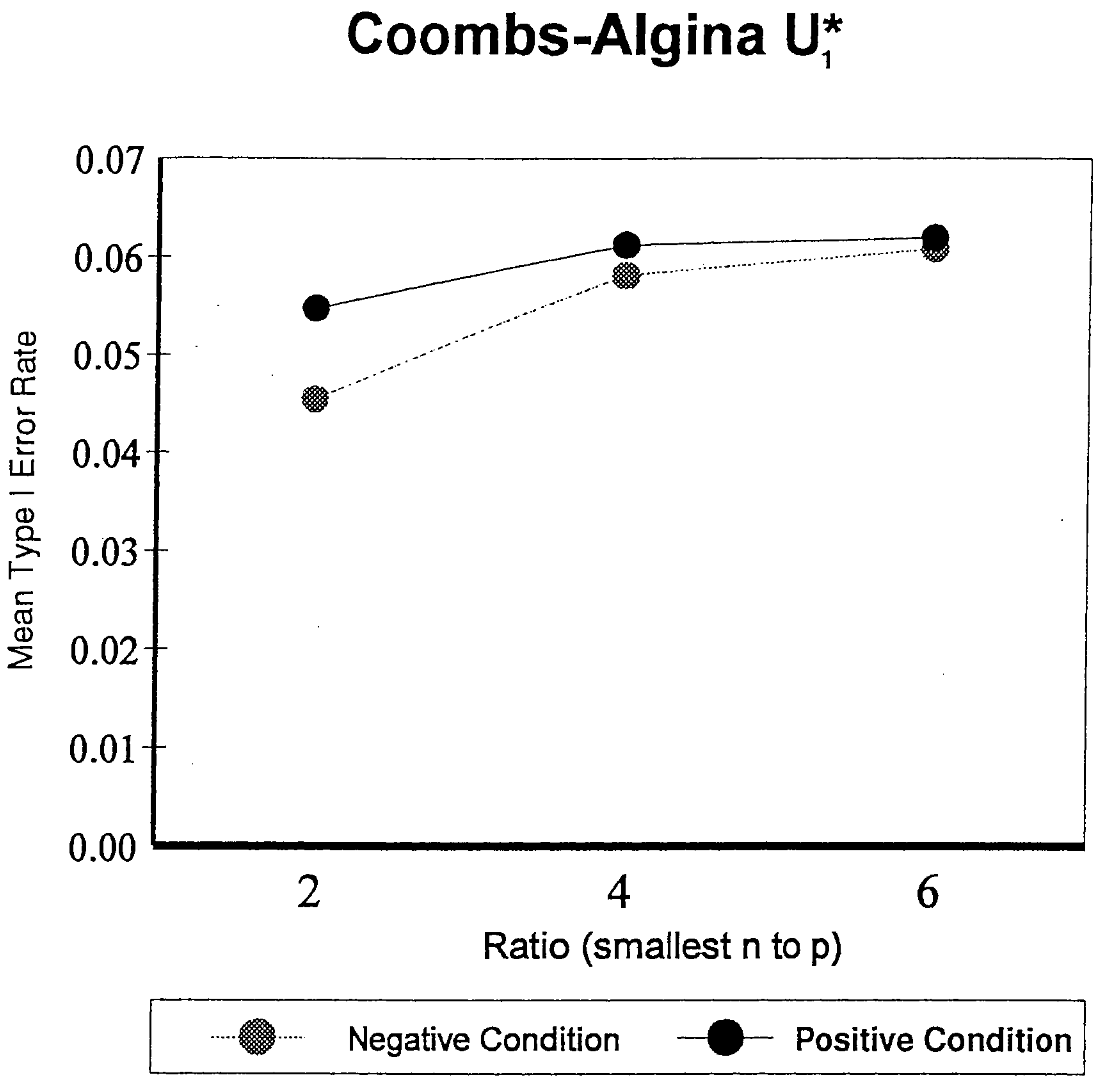


Figure 16. Estimated Type I error rates for combinations of ratios of smallest sample size to number of dependent variables (r) and relationship between sample size and covariance matrices (s) for the Coombs-Algina U_1* test.

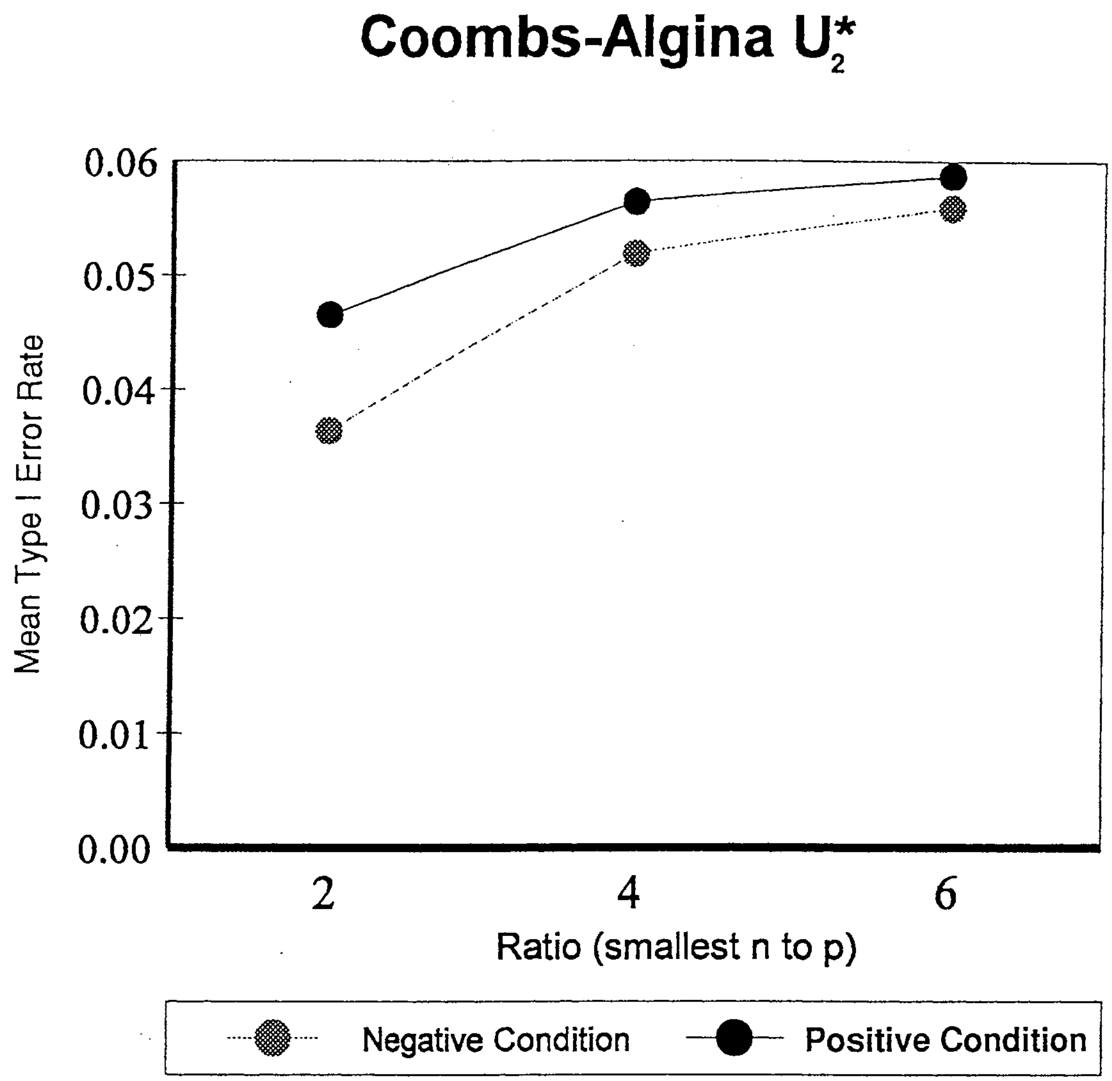


Figure 17. Estimated Type I error rates for combinations of ratios of smallest sample size to number of dependent variables (r) and relationship between sample size and covariance matrices (s) for the Coombs-Algina U_2^* test.

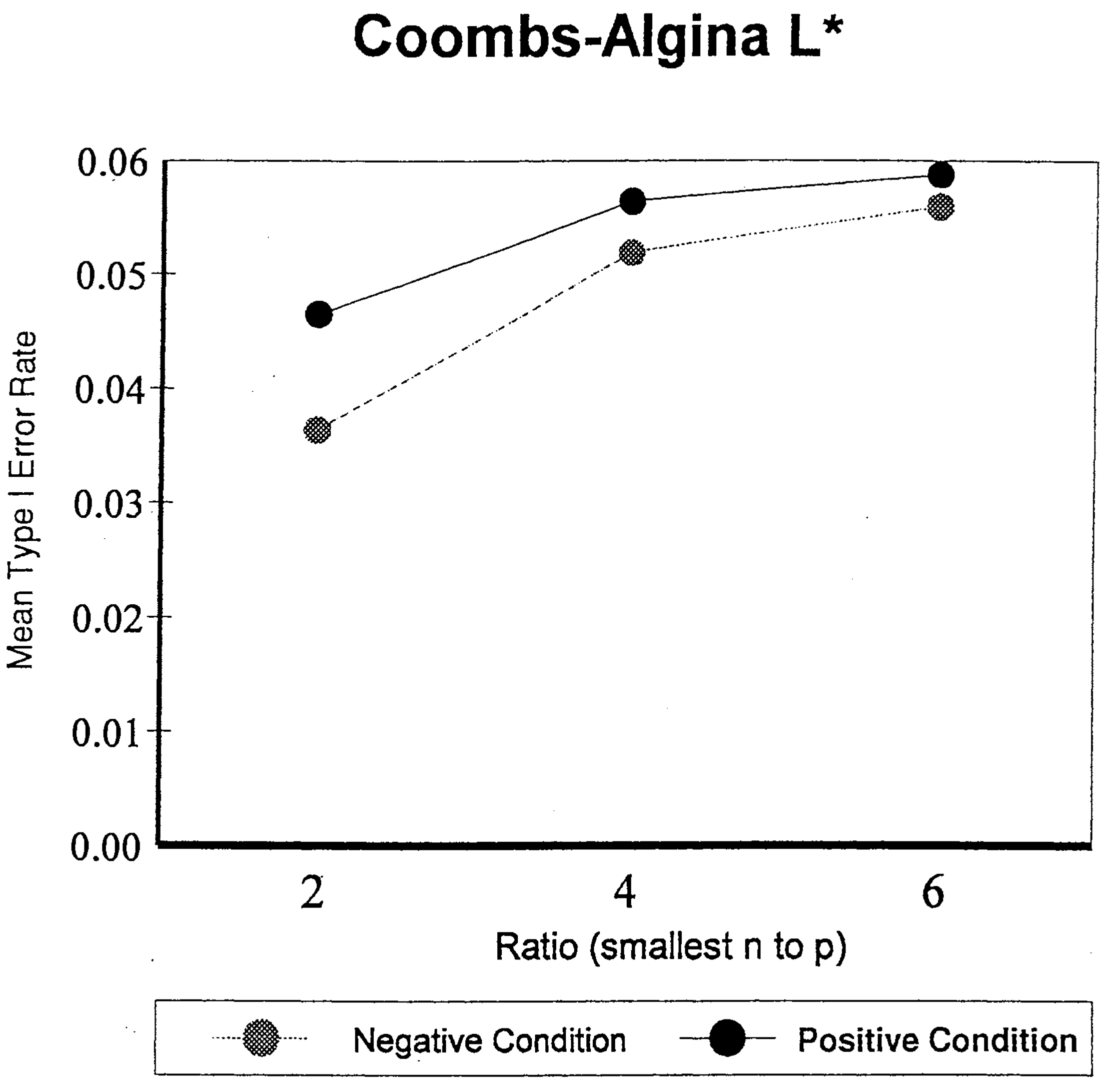


Figure 18. Estimated Type I error rates for combinations of ratios of smallest sample size to number of dependent variables (r) and relationship between sample size and covariance matrices (s) for the Coombs-Algina L* test.

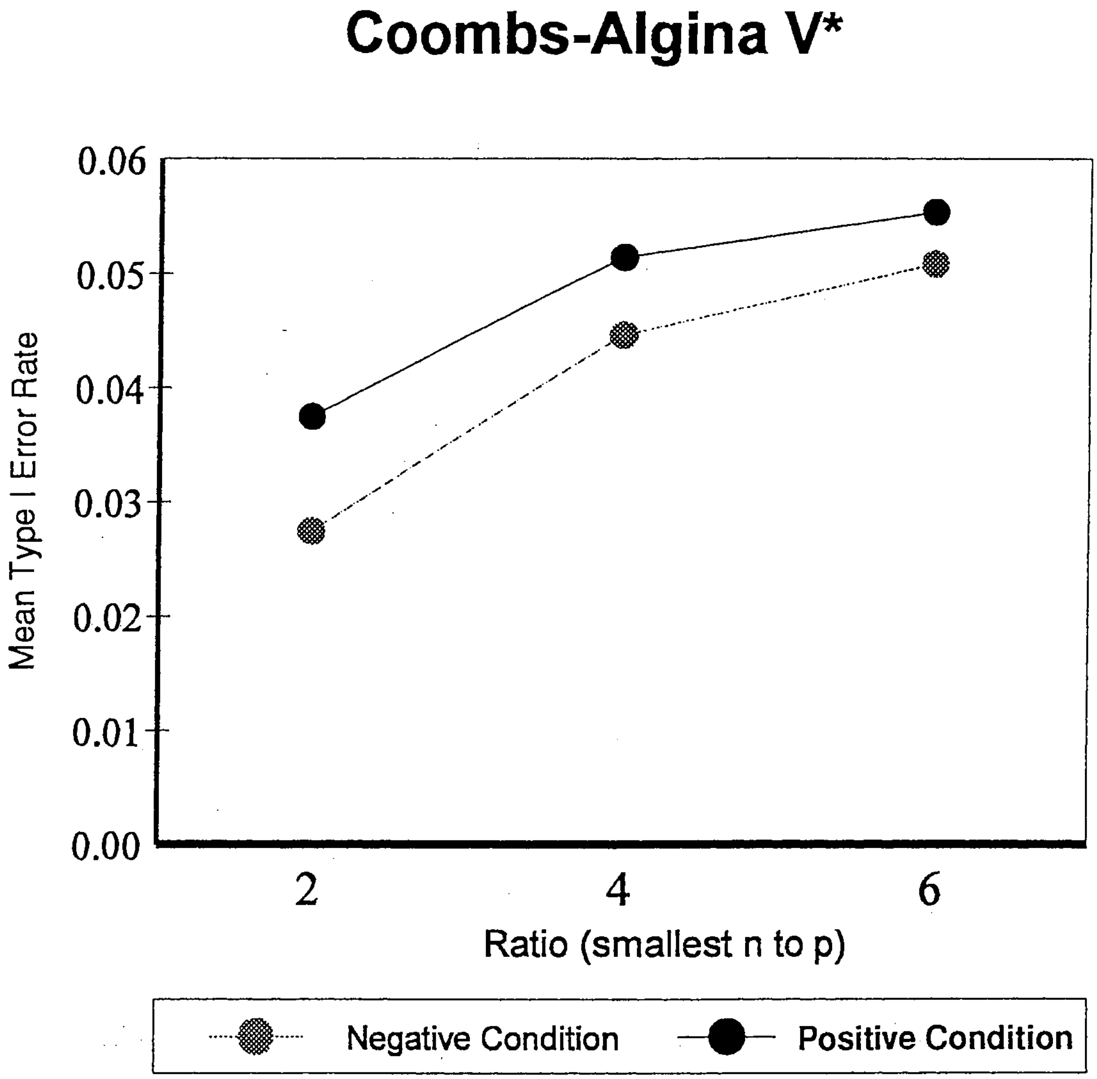
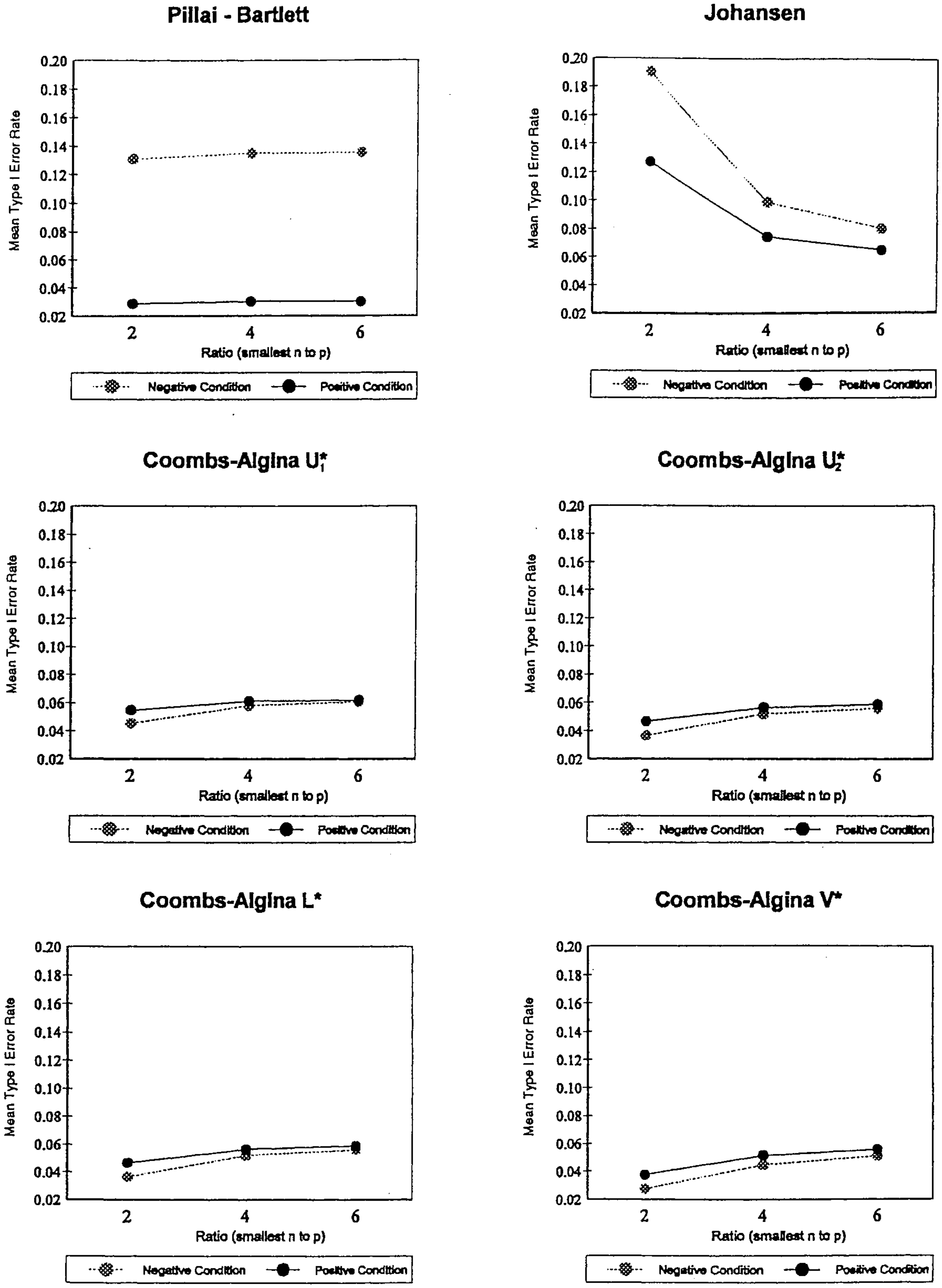


Figure 19. Estimated Type I error rates for combinations of ratios of smallest sample size to number of dependent variables (r) and relationship between sample size and covariance matrices (s) for the Coombs-Algina V* test.



Pillai - Bartlett

Johansen

Coombs-Algina U_1^*

Coombs-Algina U_2^*

Coombs-Algina L*

Coombs-Algina V*

Figure 20. Estimated Type I error rates for combinations of ratios (smallest n to p) and relationship between sample size and covariance matrices (s) for six tests.

liberal in the negative condition and when $r < 4$. Estimated mean Type I error rate in the negative condition is larger than in the positive condition. It decreases as r increases from 2 to 4 to 6, becoming adequate in the positive condition ($s = 0$) when $r \geq 4$. The differences between mean estimated Type I error rates for the positive and negative conditions decrease as r increases from 2 to 4 to 6. Estimated mean Type I error rate for the Johansen test never falls below .065, its value in the positive condition when the ratio between smallest sample size and number of dependent variables is 6.

Plots of all four Coombs-Algina tests have the same interaction pattern for s and r, which is not unexpected given that all are defined in terms of the same matrix and are generalizations of the same univariate test. The positive condition shows higher estimated mean Type I error rates with differences declining as r increases from 2 to 6. Further mean $\hat{\tau}$ increases with r with larger increases from 2 to 4 than from 4 to 6. All four Coombs-Algina tests show approximately nominal rates with mean $\hat{\tau}$ varying from .0619 (Coombs-Algina $\underline{U}_1^*$, $s = 0$, $r = 6$) to .0274 (Coombs-Algina $\underline{V}^*$, $s = 1$, $r = 2$).

Effects of Number of Groups and Ratio Between Smallest Sample Size and Number of Dependent Variables (k and r). Figures 21-26 contain the mean plots of the combinations of number of groups sampled (k) and ratio between smallest sample size and number of dependent variables (r) for the six criteria individually. Figure 27 includes them all to facilitate comparison.

The Pillai-Bartlett test is liberal with mean $\hat{\tau}$ varying from .0876 ($k = 6$, $r = 6$) to .0763 ($k = 3$, $r = 2$). The estimated mean Type I error rate tends to increase very slightly with r and is higher for $k = 6$ than for $k = 3$.

In the Johansen test mean $\hat{\tau}$ decreases as r increases from 2 to 4 to 6 and increases as k increases from 3 to 6. The differences between mean Type I error

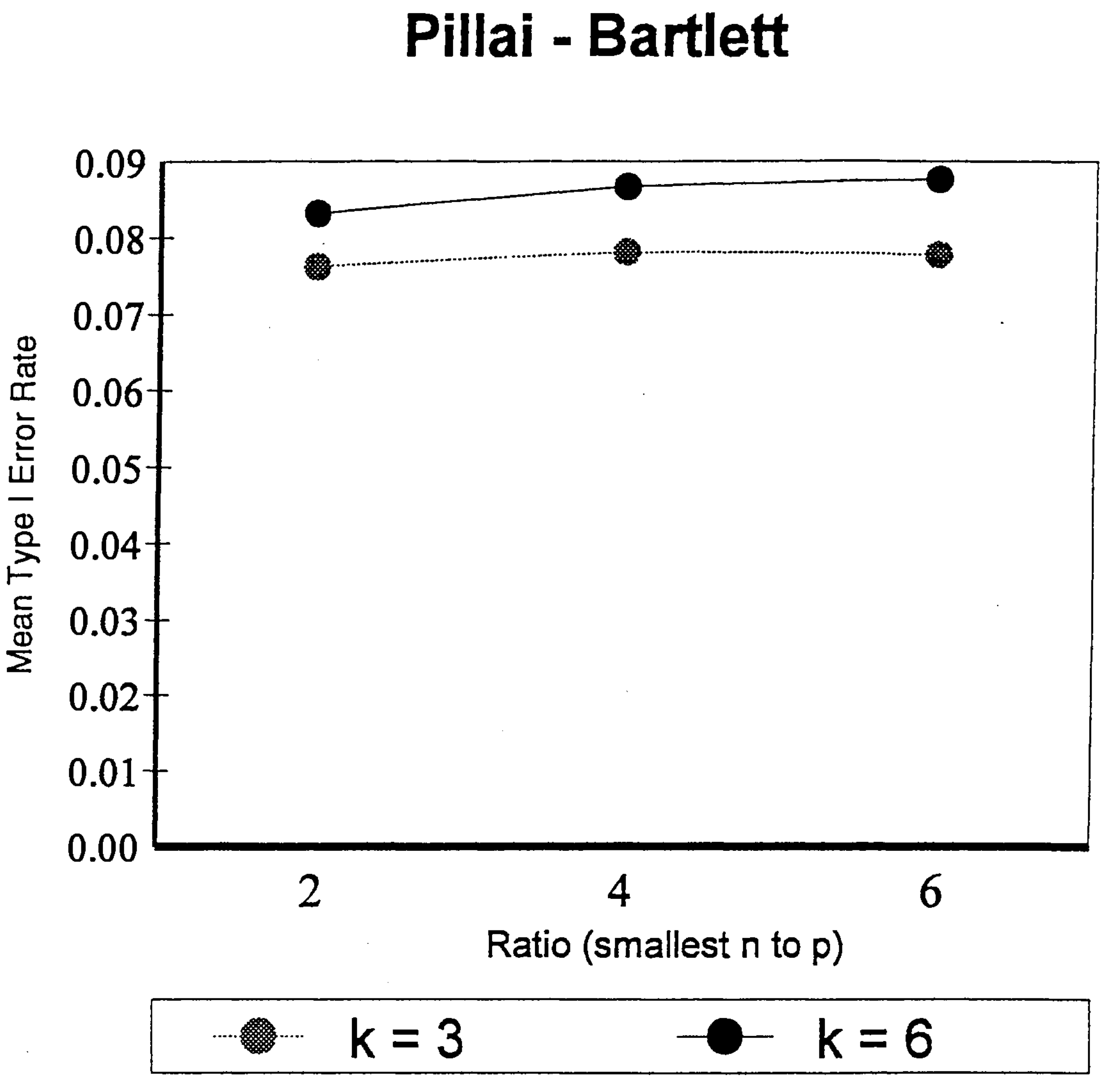


Figure 21. Estimated Type I error rates for combinations of ratesof smallest sample size to number of dependent variables (r) and number of groups (k) for the Pillai-Bartlett test.

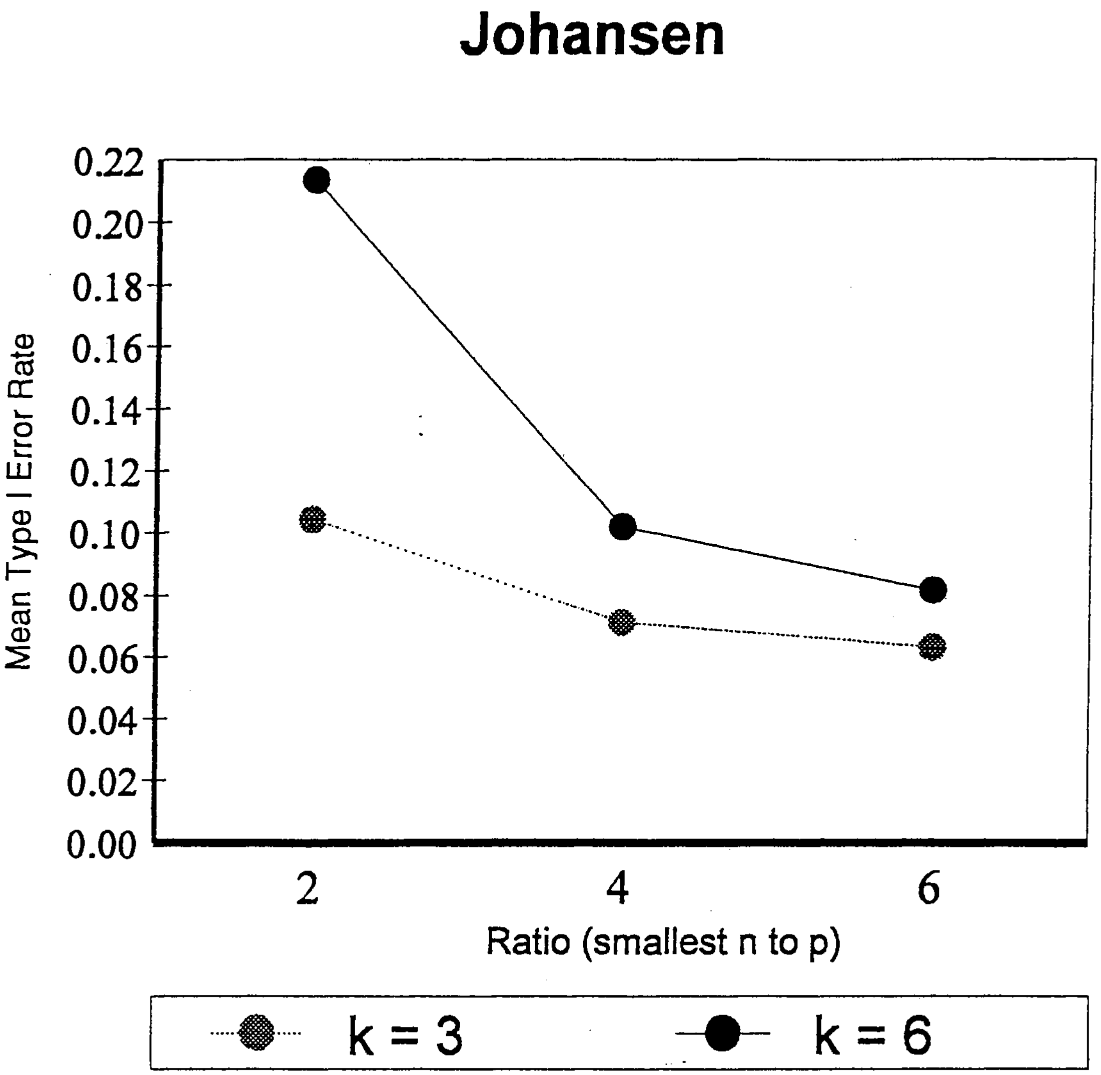


Figure 22. Estimated Type I error rates for combinations of ratesof smallest sample size to number of dependent variables (r) and number of groups (k) for the Johansen test.

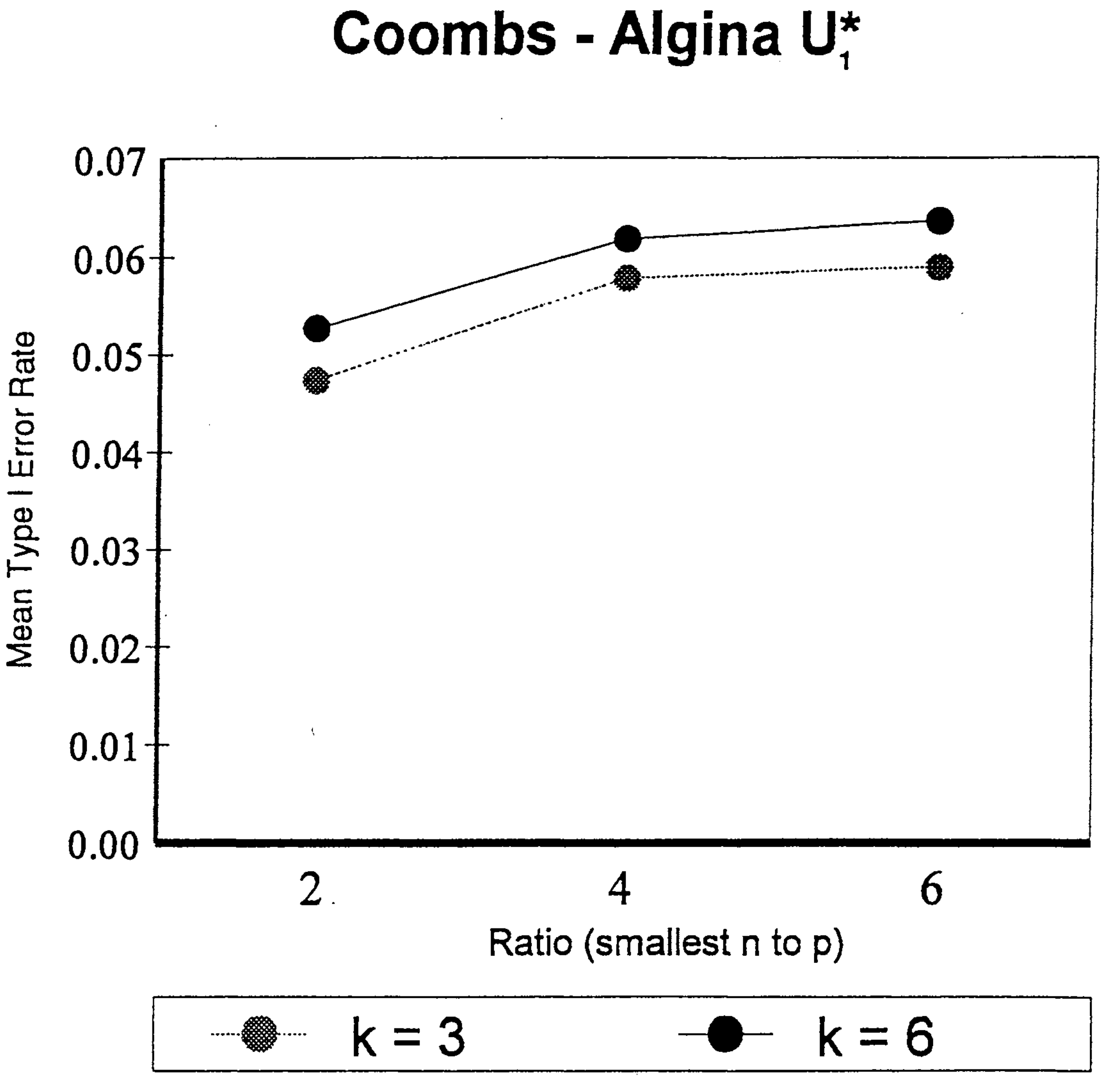


Figure 23. Estimated Type I error rates for combinations of ratesof smallest sample size to number of dependent variables (r) and number of groups (k) for the Coombs-Algina U_1* test.

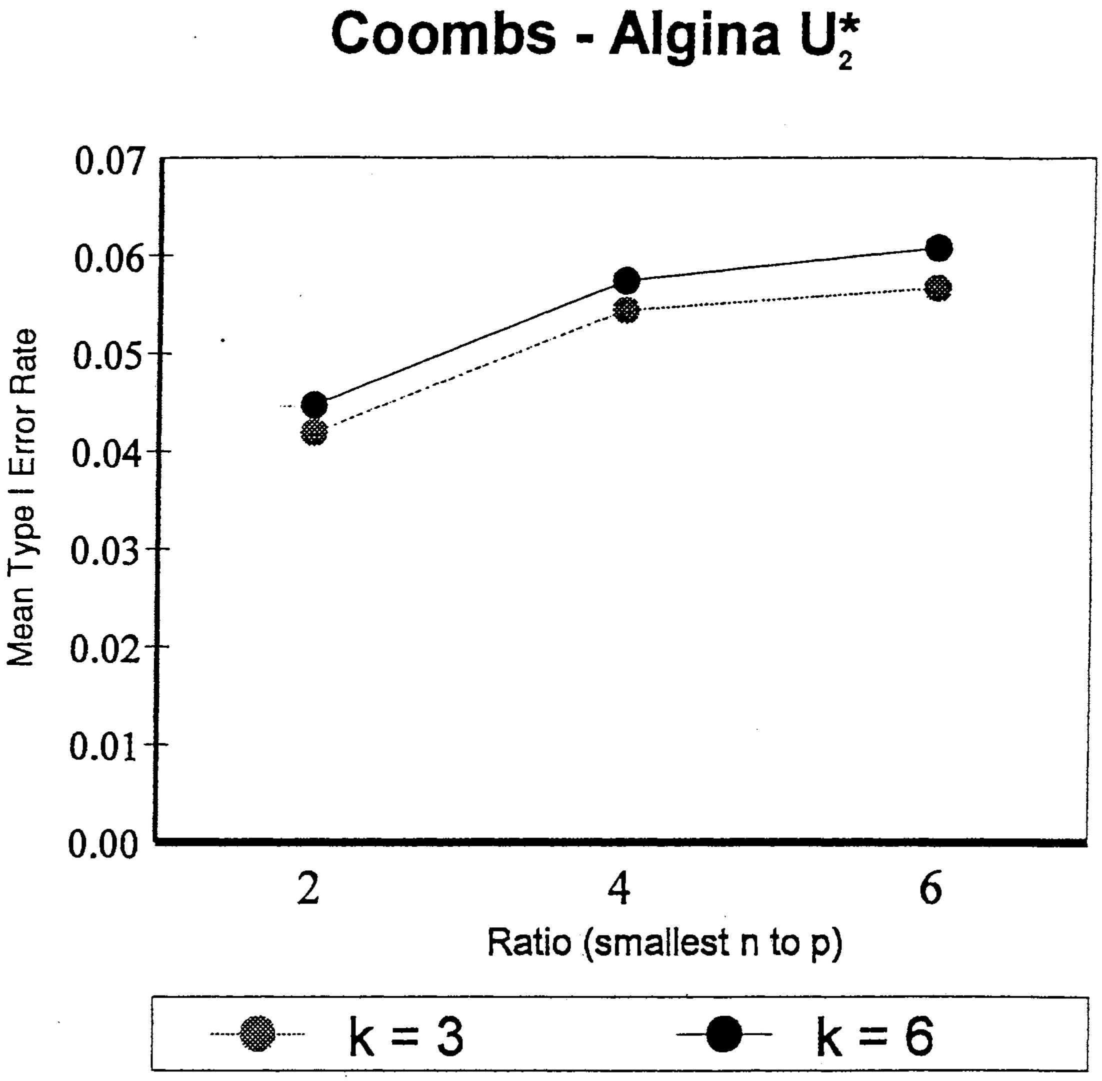


Figure 24. Estimated Type I error rates for combinations of ratesof smallest sample size to number of dependent variables (r) and number of groups (k) for the Coombs-Algina U_2^* test.

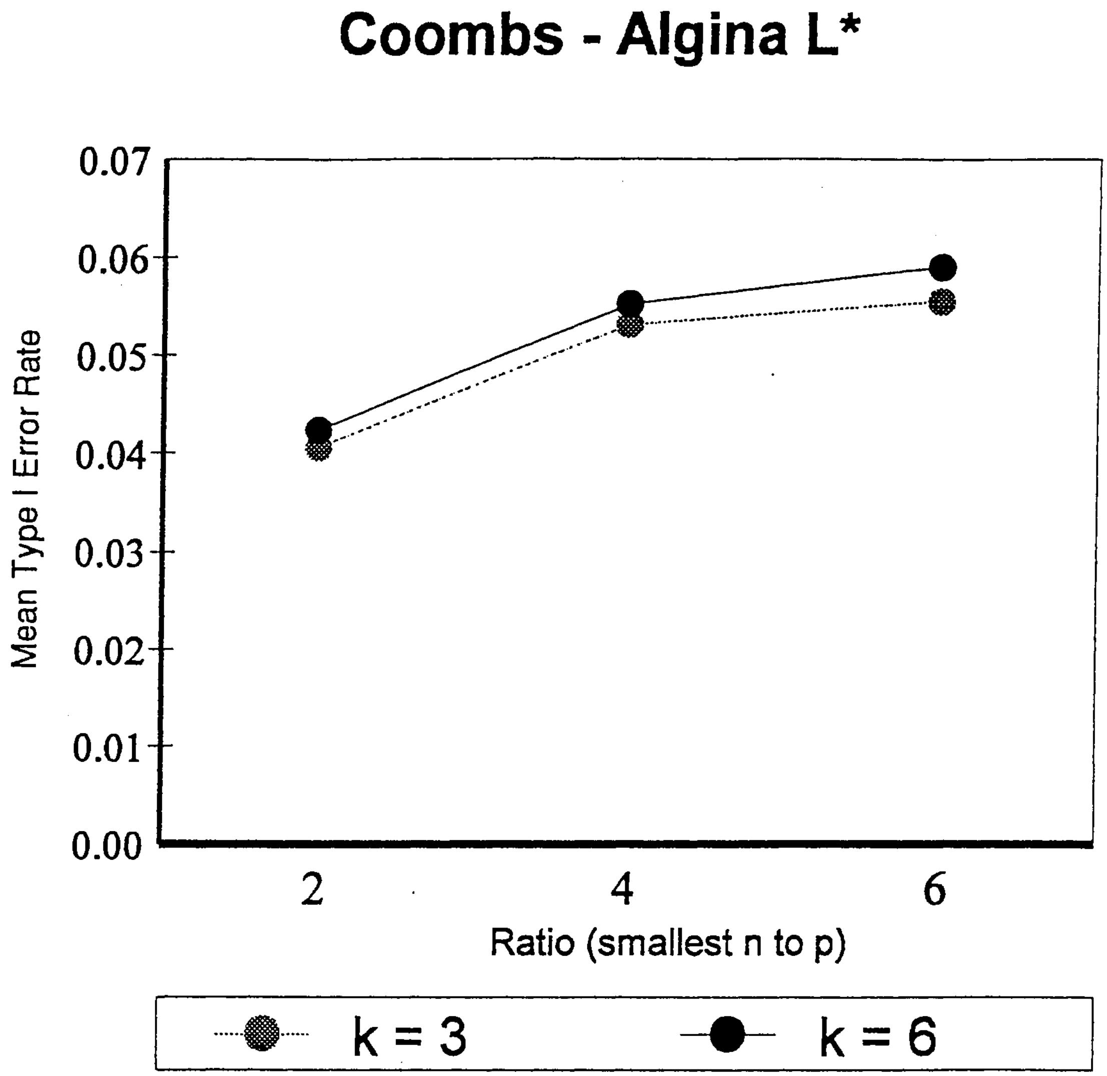


Figure 25. Estimated Type I error rates for combinations of ratesof smallest sample size to number of dependent variables (r) and number of groups (k) for the Coombs-Algina L* test.

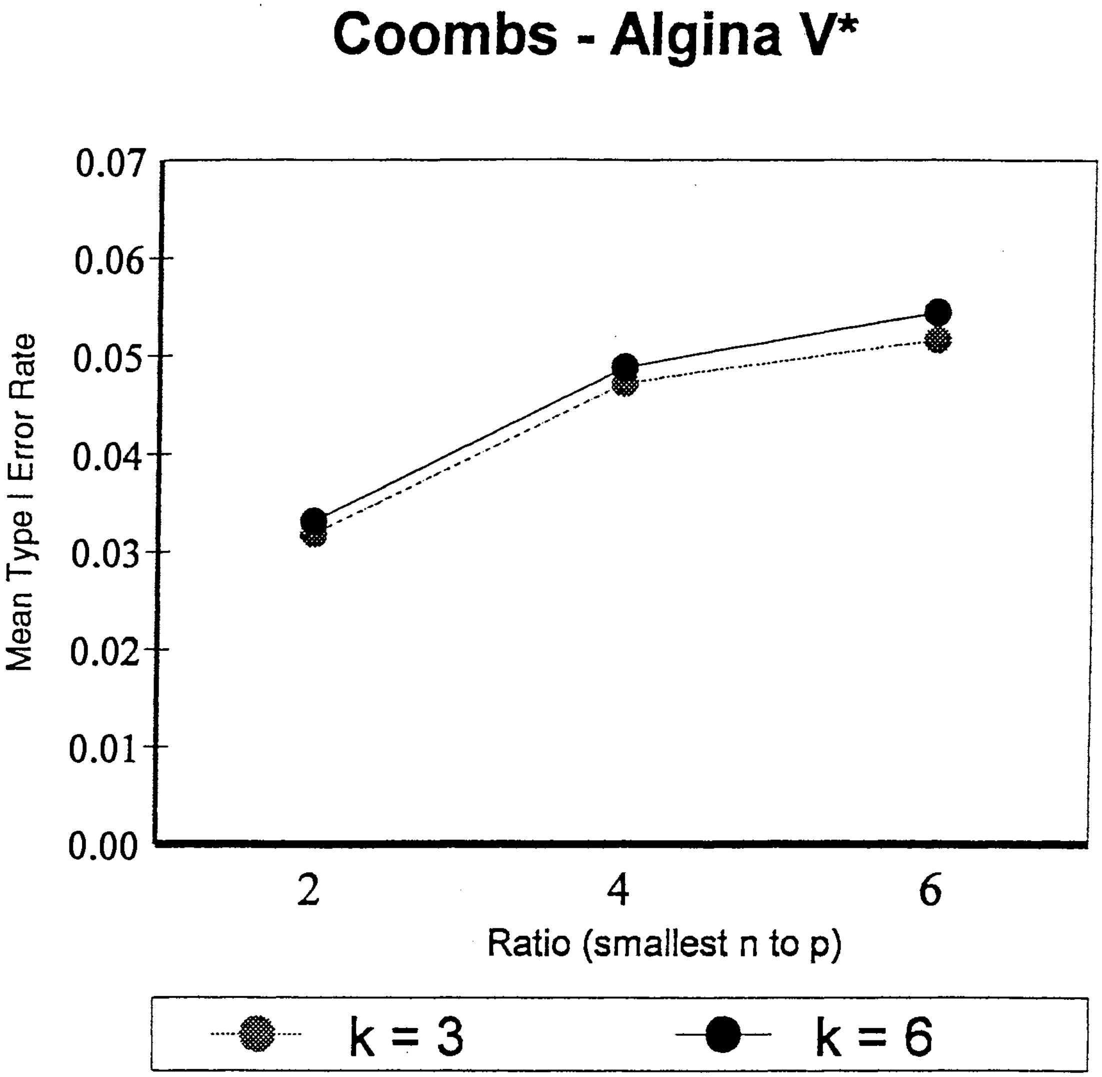


Figure 26. Estimated Type I error rates for combinations of ratesof smallest sample size to number of dependent variables (r) and number of groups (k) for the Coombs-Algina V* test.

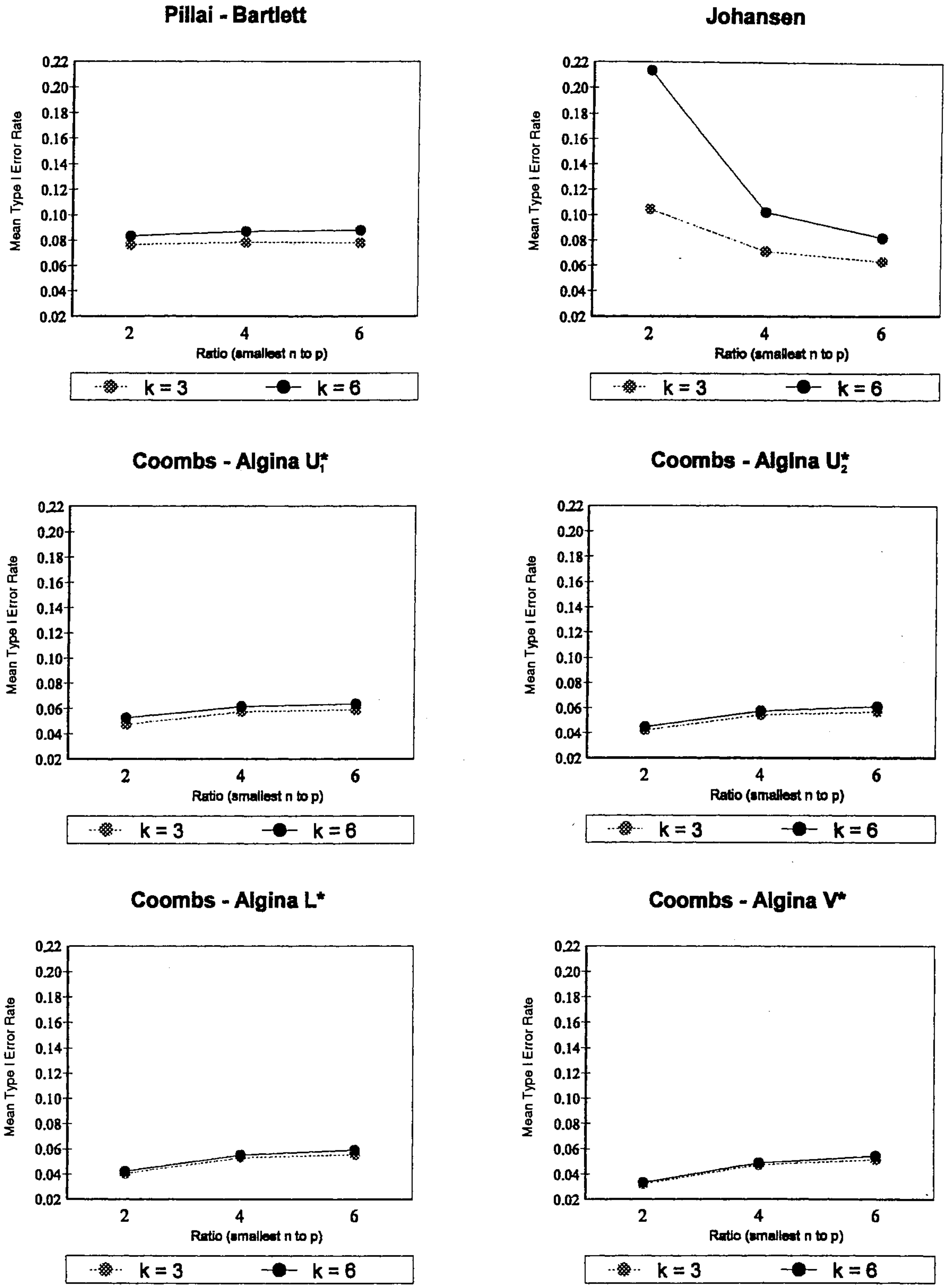


Figure 27. Estimated Type I error rates for combinations of ratios (smallest $\underline{n}$ to $\underline{p}$) and number of groups ($\underline{k}$) for six tests.

rates for $\underline{k} = 6$ and $\underline{k} = 3$ decreases as $\underline{r}$ increases from 2 to 4 to 6. The Johansen test is considerably more liberal than the Pillai-Bartlett test with mean $\hat{\tau}$ varying from .0630 ($\underline{k} = 3$, $\underline{r} = 6$) to .2135 ($\underline{k} = 3$, $\underline{r} = 2$). It adequately controls Type I error rate when $\underline{k} < 6$ and $\underline{r} \geq 4$.

Estimated mean Type I error rates for the four Coombs-Algina tests are close to nominal, varying from .0637 ($\underline{U}_1^*$, $\underline{k} = 6$, $\underline{r} = 6$) to .0318 ($\underline{V}^*$, $\underline{k} = 3$, $\underline{r} = 2$). In all four tests mean $\hat{\tau}$ increases as $\underline{r}$ increases from 2 to 4 to 6. Also in all four tests mean $\hat{\tau}$ increases as $\underline{k}$ increases from 3 to 6 with differences increasing slightly as $\underline{r}$ increases. These similar results for the Coombs-Algina tests are again expected, given the nature of the criteria's definitions (all in terms of functions of the same matrix).

For all six tests mean $\hat{\tau}$ was higher for $\underline{k} = 6$ than for $\underline{k} = 3$.

Effect of Number of Groups and Distribution Type ($\underline{k}$ and $\underline{DT}$). The mean plots involving the combinations of number of groups sampled and distribution type appear individually in Figures 28-33 and collectively in Figure 34. They show the Pillai-Bartlett and Johansen tests to be liberal in all cases, the Johansen test more so except when sampling from a small ($\underline{k} = 3$) number of normal distributions. All four Coombs-Algina tests had estimated mean Type I error rates close to nominal levels.

The Pillai-Bartlett test and the four Coombs-Algina tests all had higher estimated mean Type I error rates for the normal distribution when $\underline{k} = 6$ and for the exponential distribution when $\underline{k} = 3$. In the Johansen test, mean $\hat{\tau}$ was higher in the exponential distribution than in the normal distribution for both $\underline{k} = 3$ and $\underline{k} = 6$, although more pronounced for $\underline{k} = 6$.

The Coombs-Algina $\underline{V}^*$ test was slightly conservative in all combinations; $\underline{U}_1^*$ was slightly liberal in all combinations; and $\underline{U}_2^*$ and $\underline{L}^*$ were slightly liberal

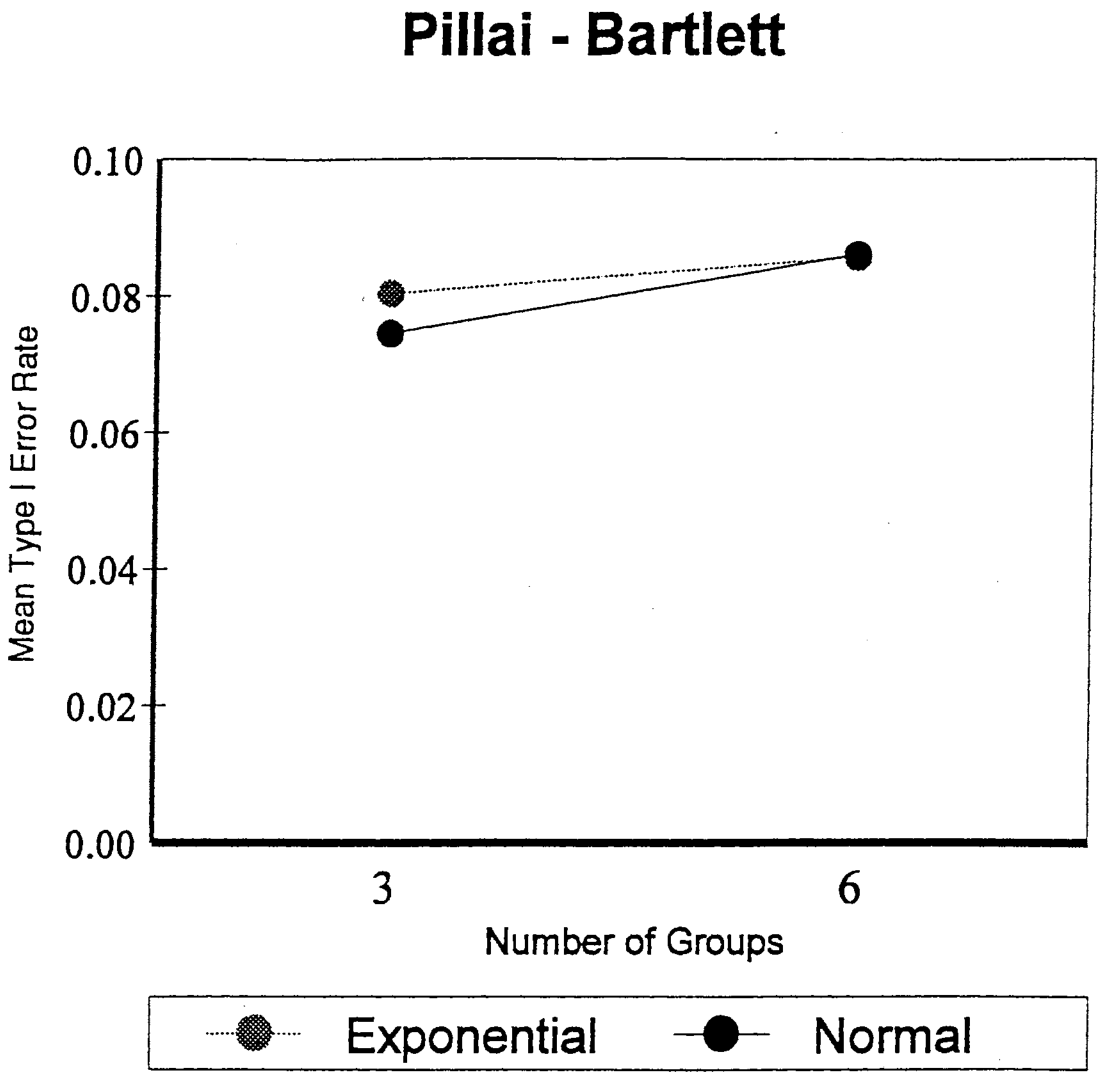


Figure 28. Estimated Type I error rates for combinations of numbers of groups (k) and ditribution types (DT) for the Pillai-Bartlett test.

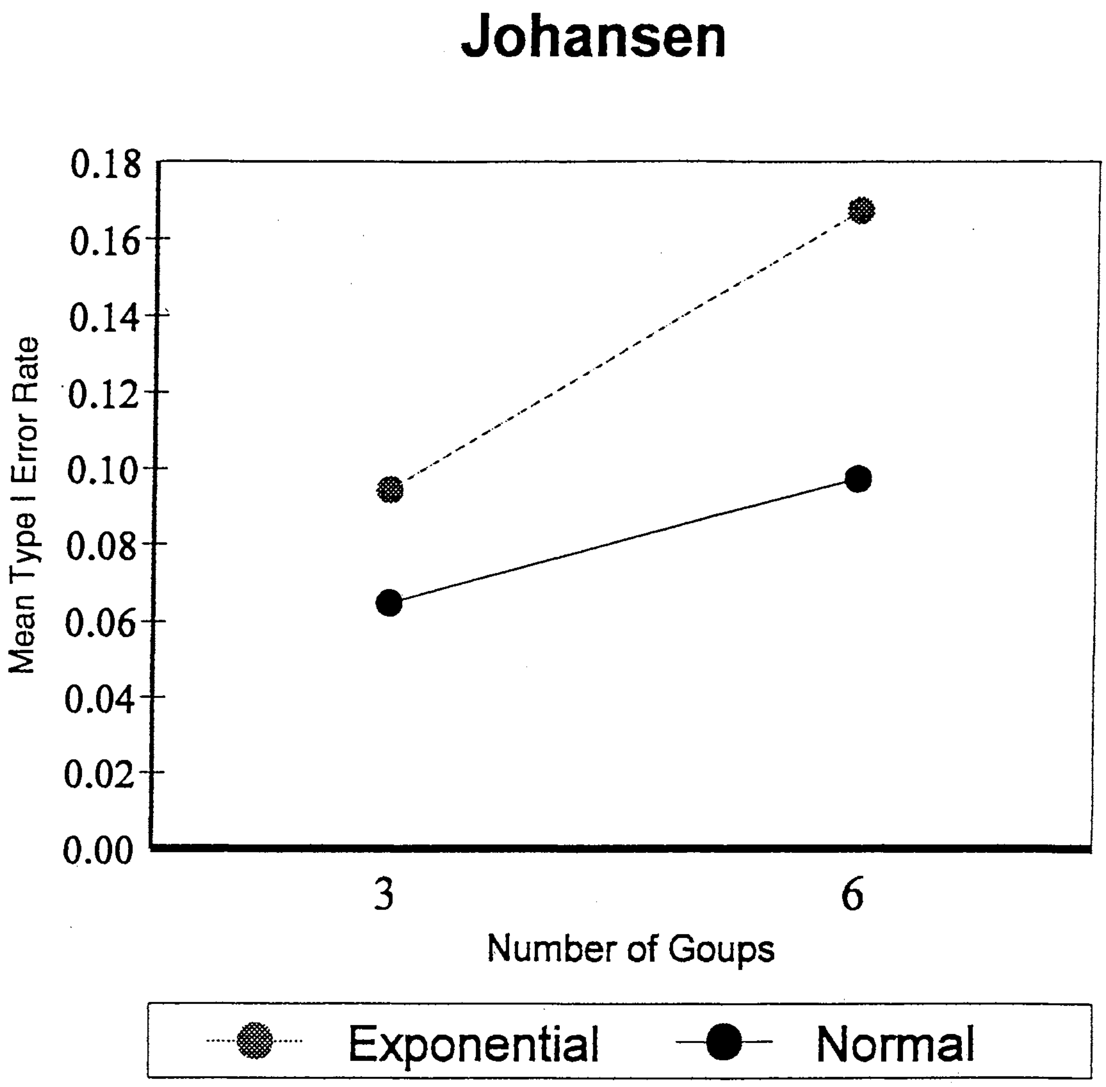


Figure 29. Estimated Type I error rates for combinations of numbers of groups (k) and ditribution types (DT) for the Johansen test.

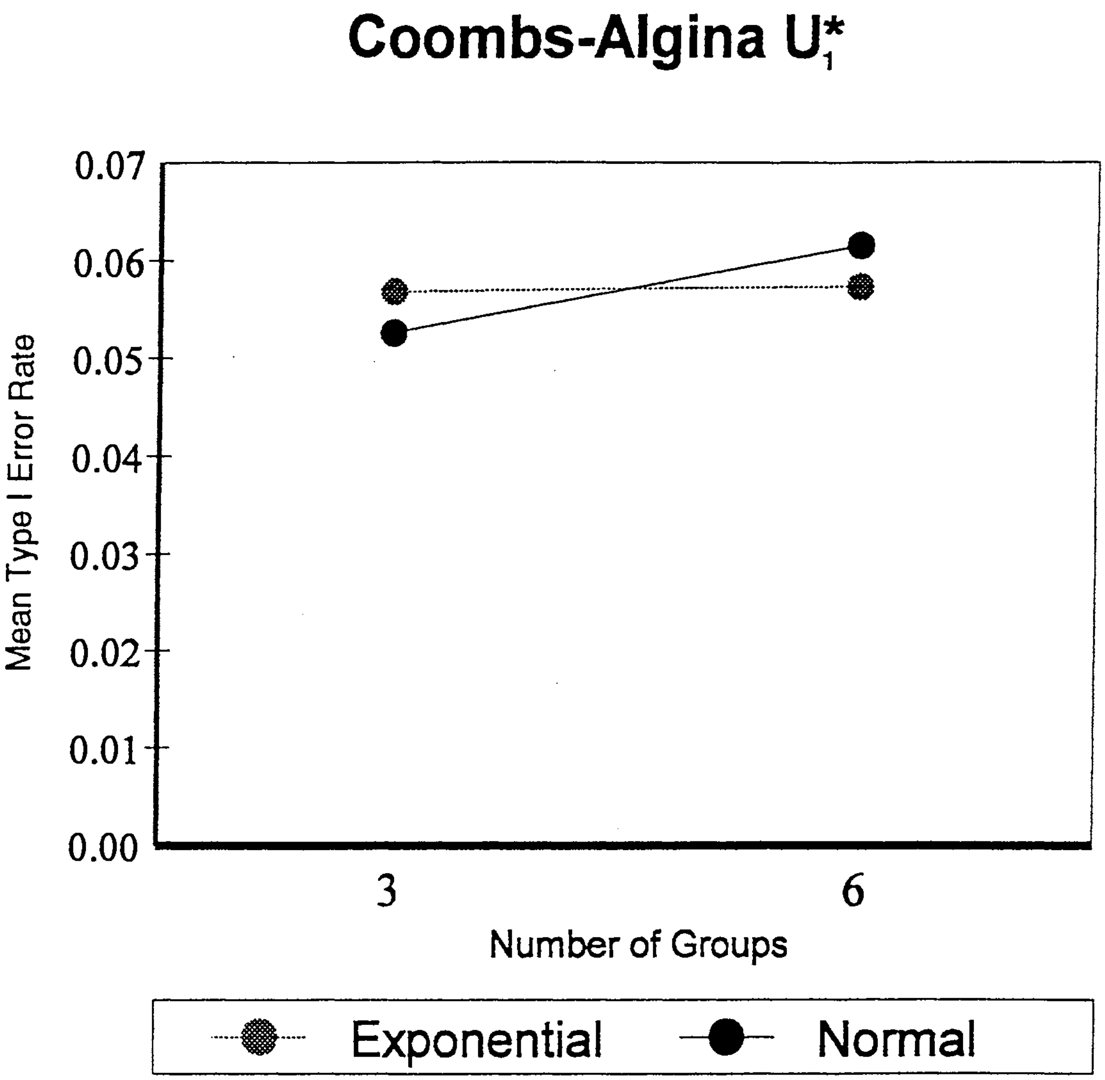


Figure 30. Estimated Type I error rates for combinations of numbers of groups (k) and ditribution types (DT) for the Coombs-Algina U_1^* test.

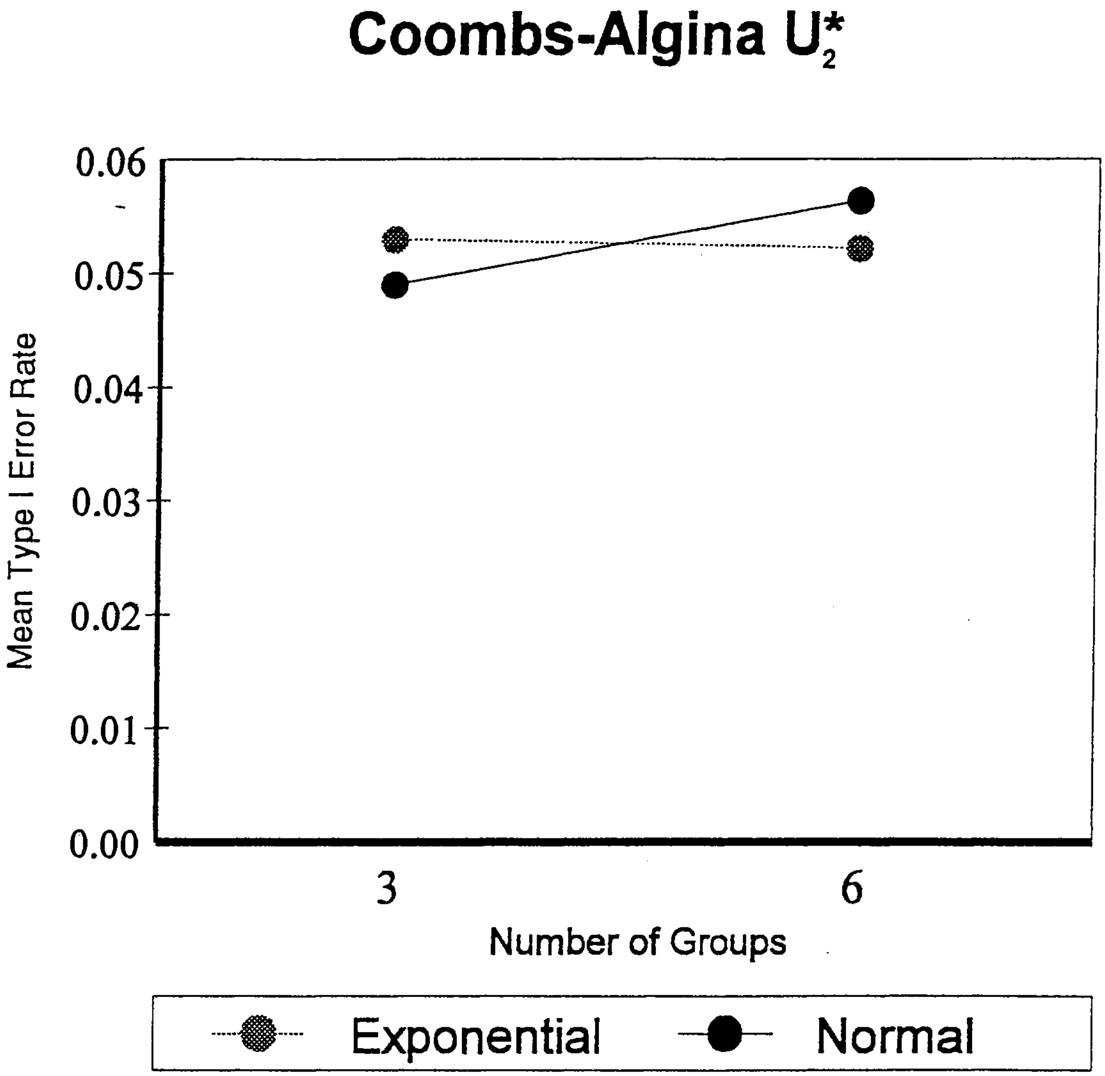


Figure 31. Estimated Type I error rates for combinations of numbers of groups (k) and ditribution types (DT) for the Coombs-Algina U_2* test.

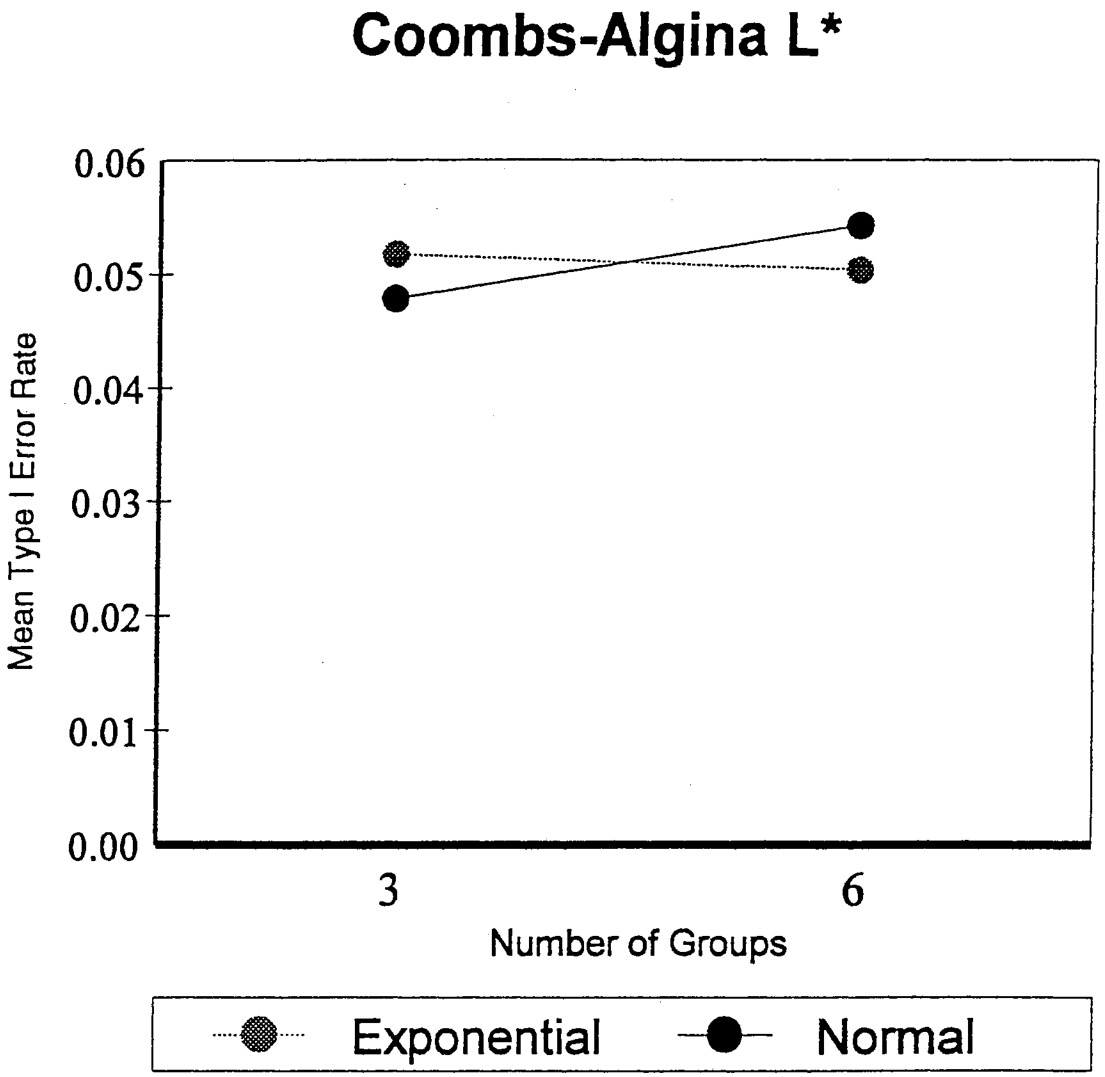


Figure 32. Estimated Type I error rates for combinations of numbers of groups (k) and ditribution types (DT) for the Coombs-Algina L* test.

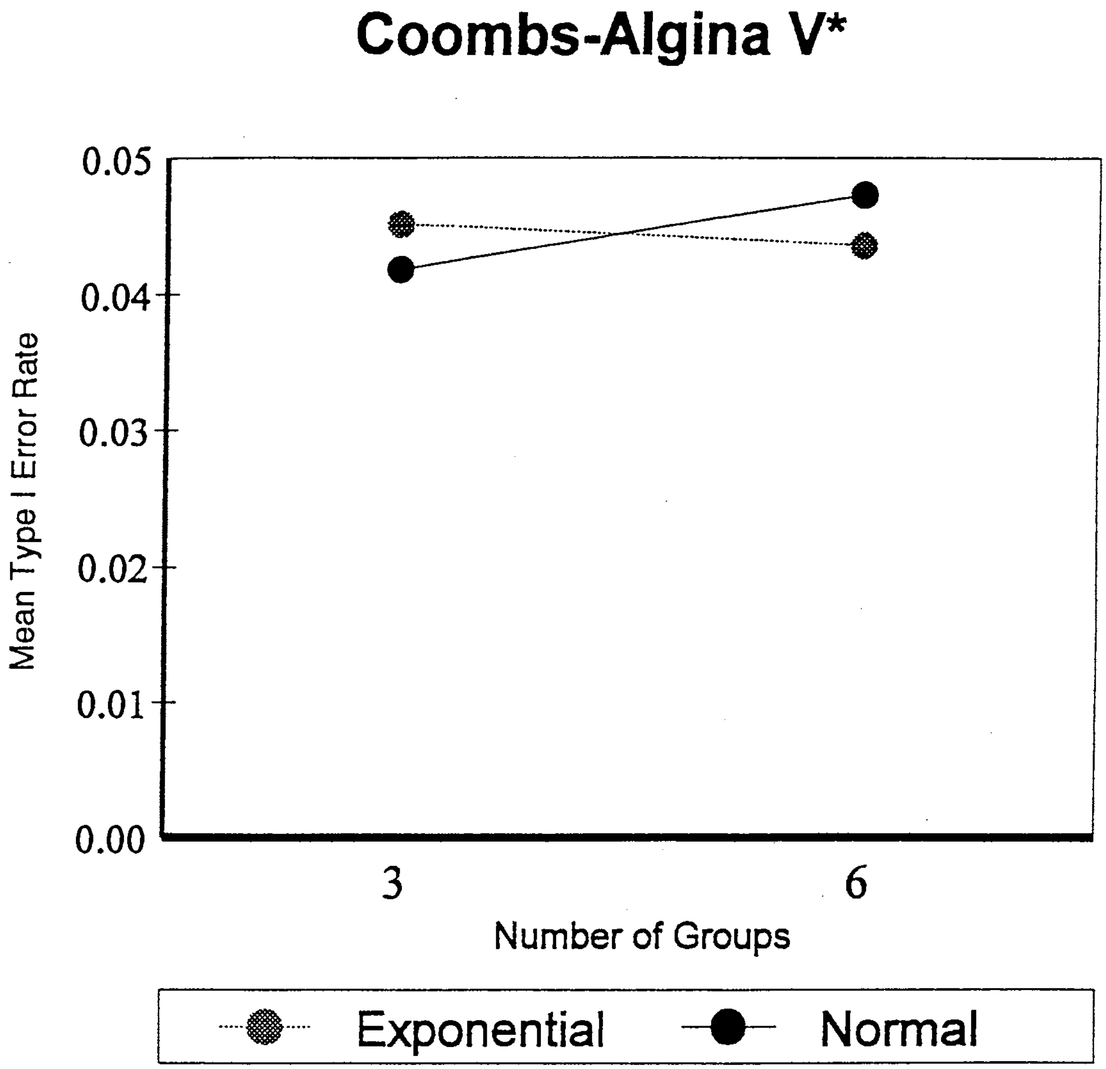
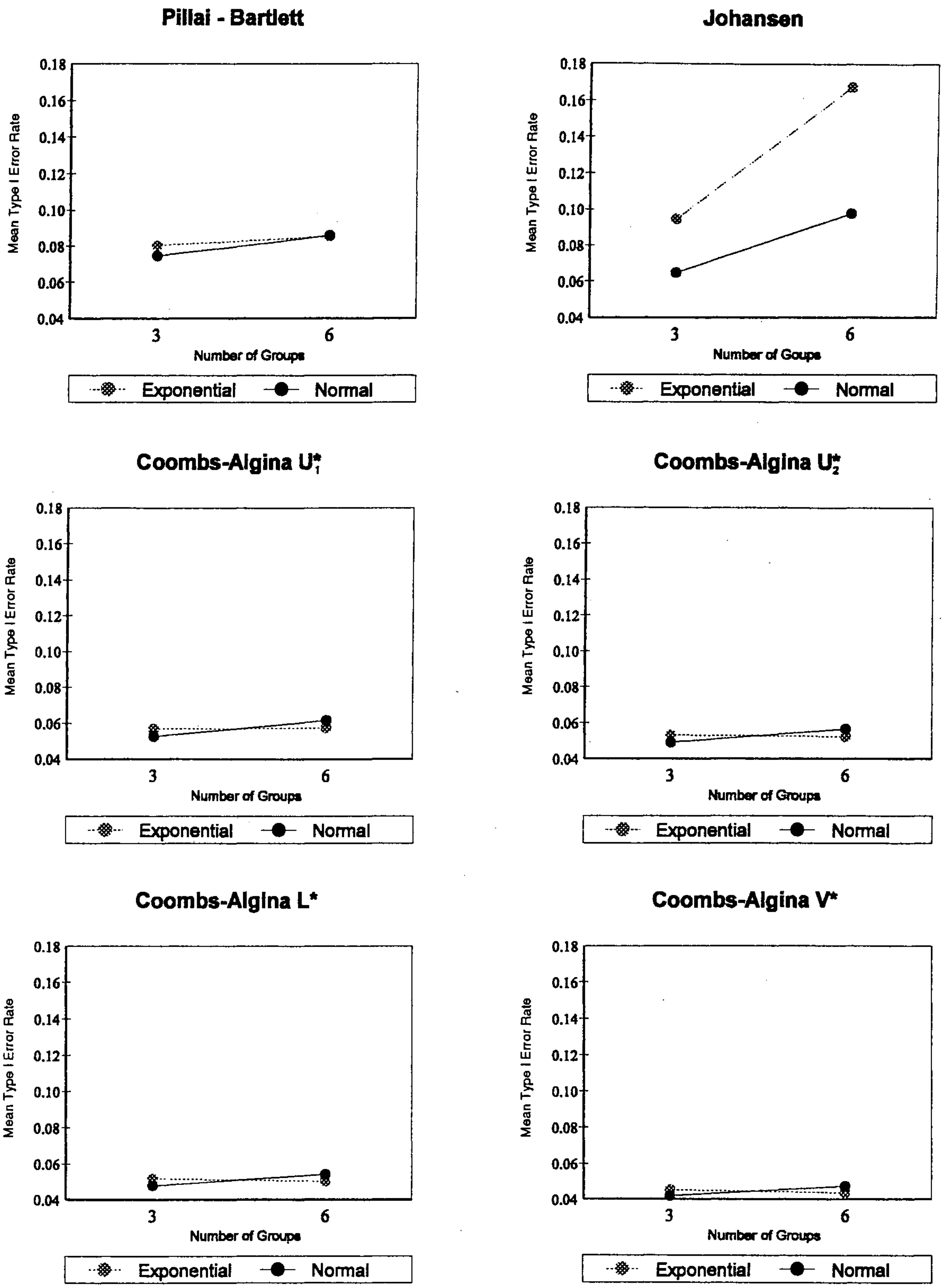


Figure 33. Estimated Type I error rates for combinations of numbers of groups (k) and ditribution types (DT) for the Coombs-Algina V* test.



Pillai - Bartlett

Johansen

Coombs-Algina U_1^*

Coombs-Algina U_2^*

Coombs-Algina L*

Coombs-Algina V*

Figure 34. Estimated Type I error rates for combinations of number of groups (k) and distribution type (DT) for six tests.

except when $k = 3$ and distribution type was normal.

Effect of Number of Dependent Variables (p). The factor p, number of dependent variables, was the only factor that did not account for a practically significant proportion of variance as either a main effect or in combination with any other factor or factors. As a main effect, it accounted for only .3611% of total variance.

Power Results

Figures 35-40 depict the distributions of the estimated power levels for the Pillai-Bartlett, Johansen, Coombs-Algina U_1^*, Coombs-Algina U_2^*, Coombs-Algina L^*, and Coombs-Algina V^* tests. Table 10 reports percentiles for each of these six tests. It shows that half the powers of the Pillai-Bartlett, Johansen, Coombs-Algina U_1^*, Coombs-Algina U_2^*, Coombs-Algina L^*, and Coombs-Algina V^* tests are respectively .2190, .2438, .1016, .0932, .0868, and .0730 or less. Both

Table 10

Percentiles for Estimated Power Level

Test Criterion	Percentile				
	0	25	50	75	100
Pillai-Bartlett	.0114	.0531	.2190	.6909	.9972
Johansen	.0550	.1053	.2438	.7238	.9956
Coombs-Algina U_1^*	.0264	.0540	.1016	.6726	.9975
Coombs-Algina U_2^*	.0183	.0505	.0932	.6609	.9975
Coombs-Algina L^*	.0207	.0467	.0868	.6390	.9973
Coombs-Algina V^*	.0161	.0449	.0730	.5885	.9964

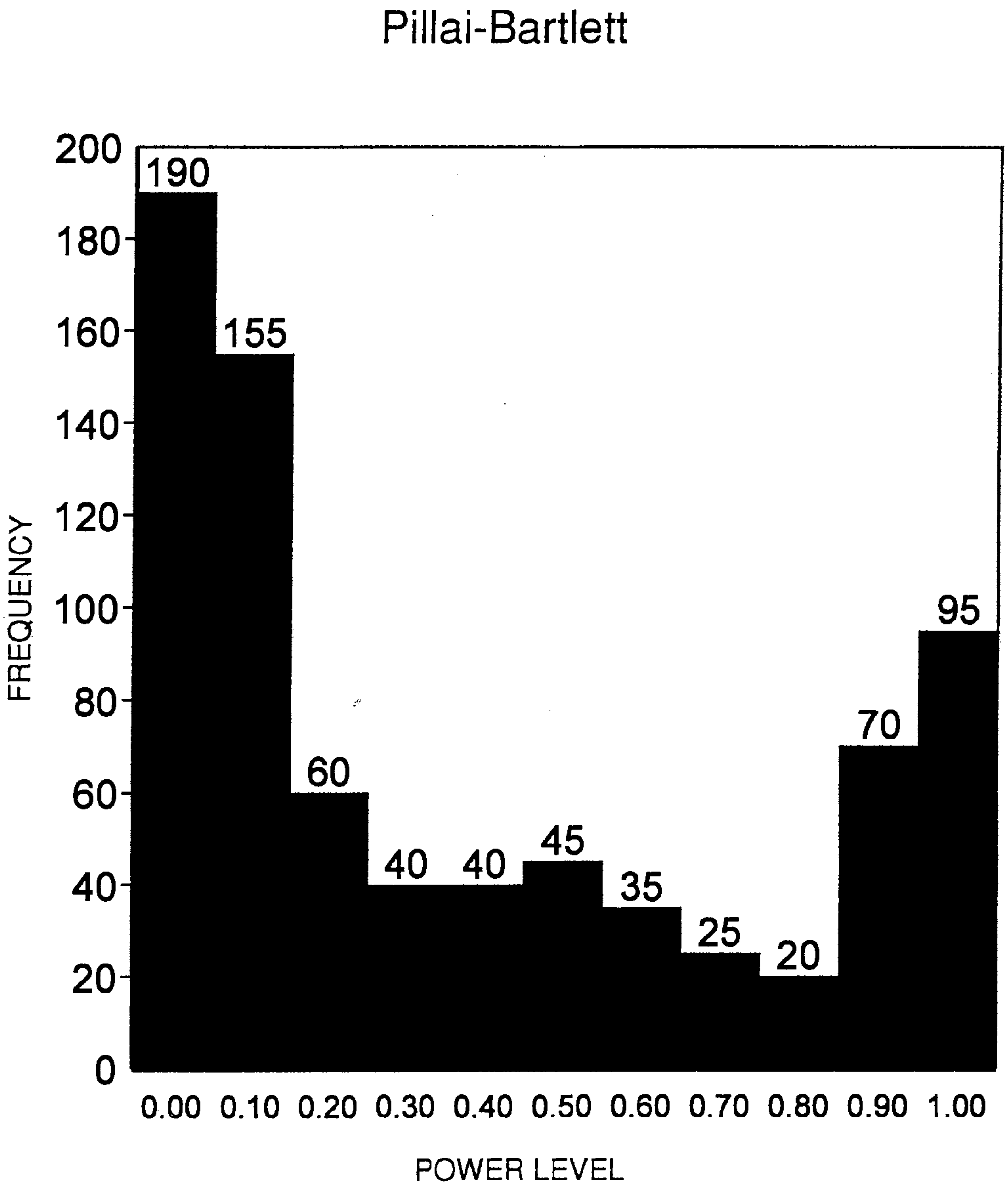


Figure 35. Frequency histogram of estimated power levels for the Pillai-Bartlett test.

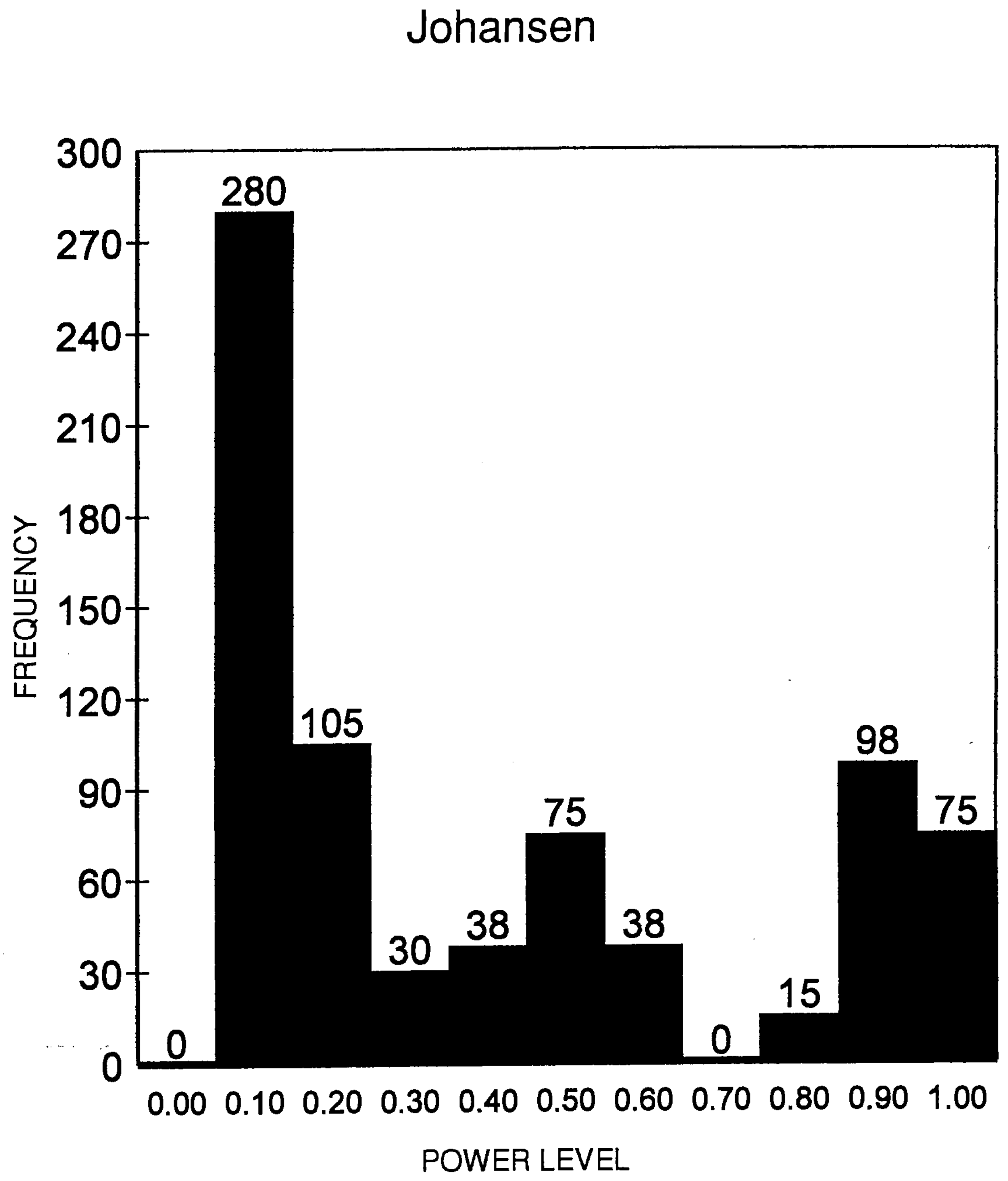


Figure 36. Frequency histogram of estimated power levels for the Johansen test.

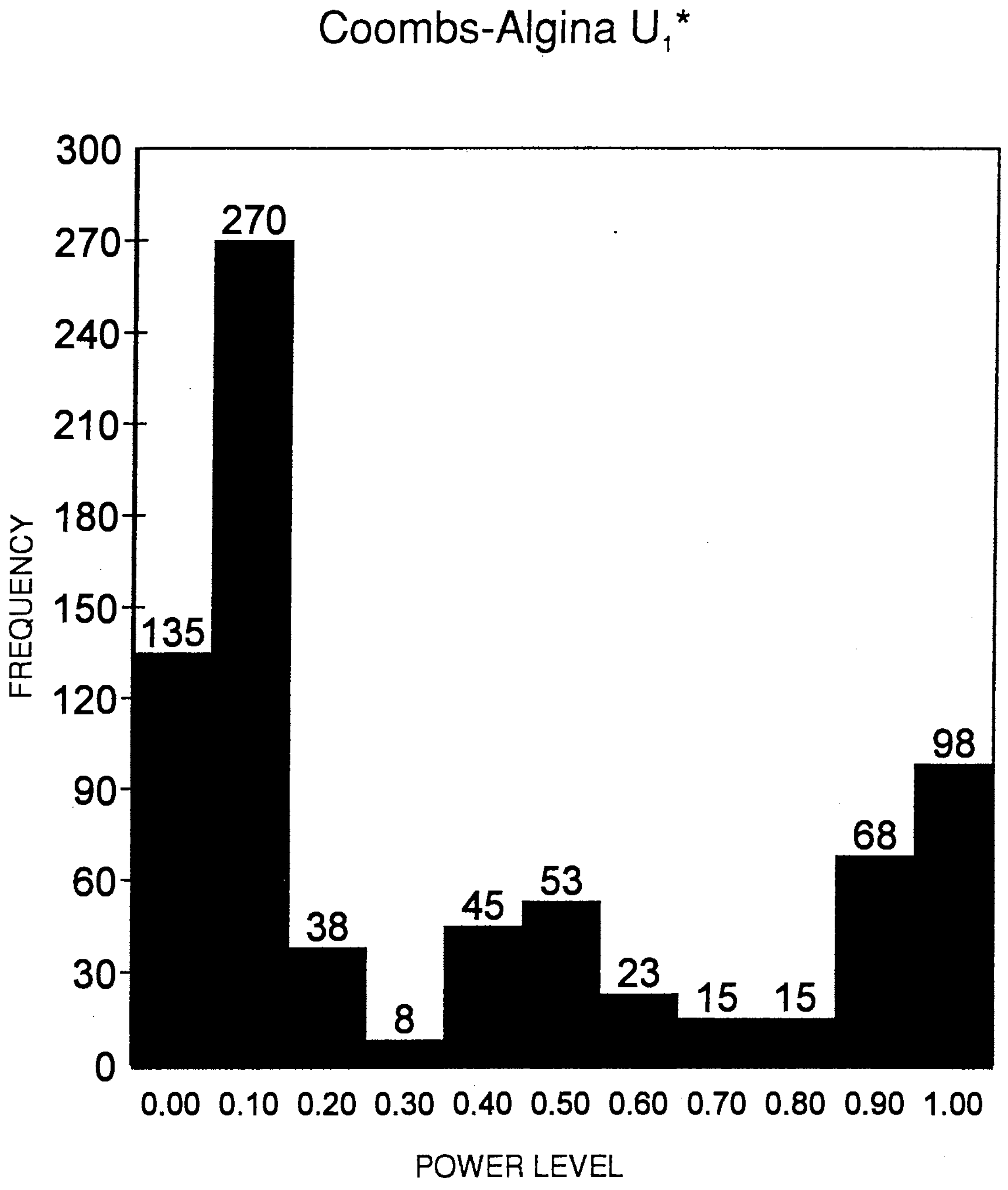


Figure 37. Frequency histogram of estimated power levels for the Coombs-Algina U_1* test.

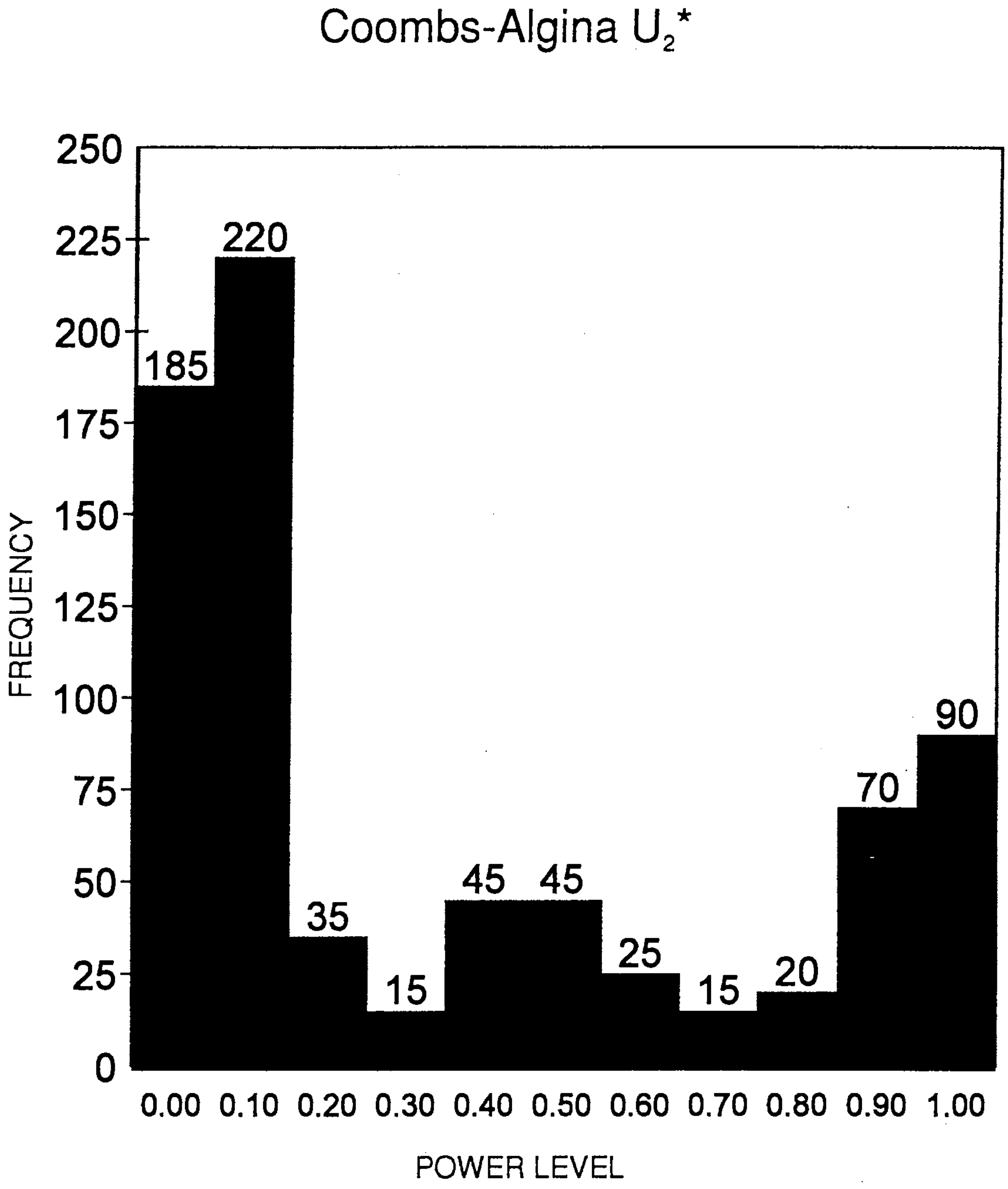


Figure 38. Frequency histogram of estimated power levels for the Coombs-Algina U_2* test.

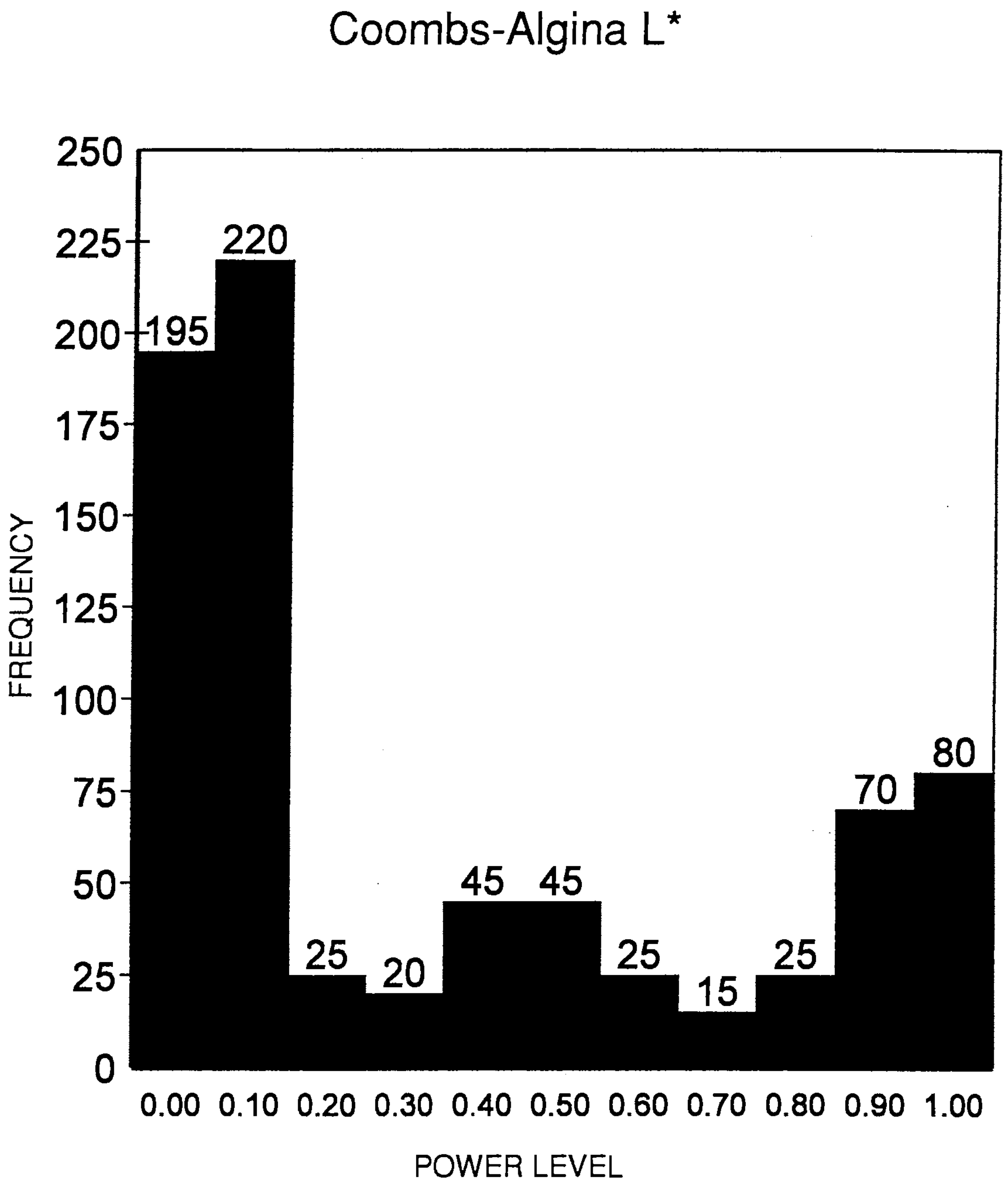


Figure 39. Frequency histogram of estimated power levels for the Coombs-Algina L* test.

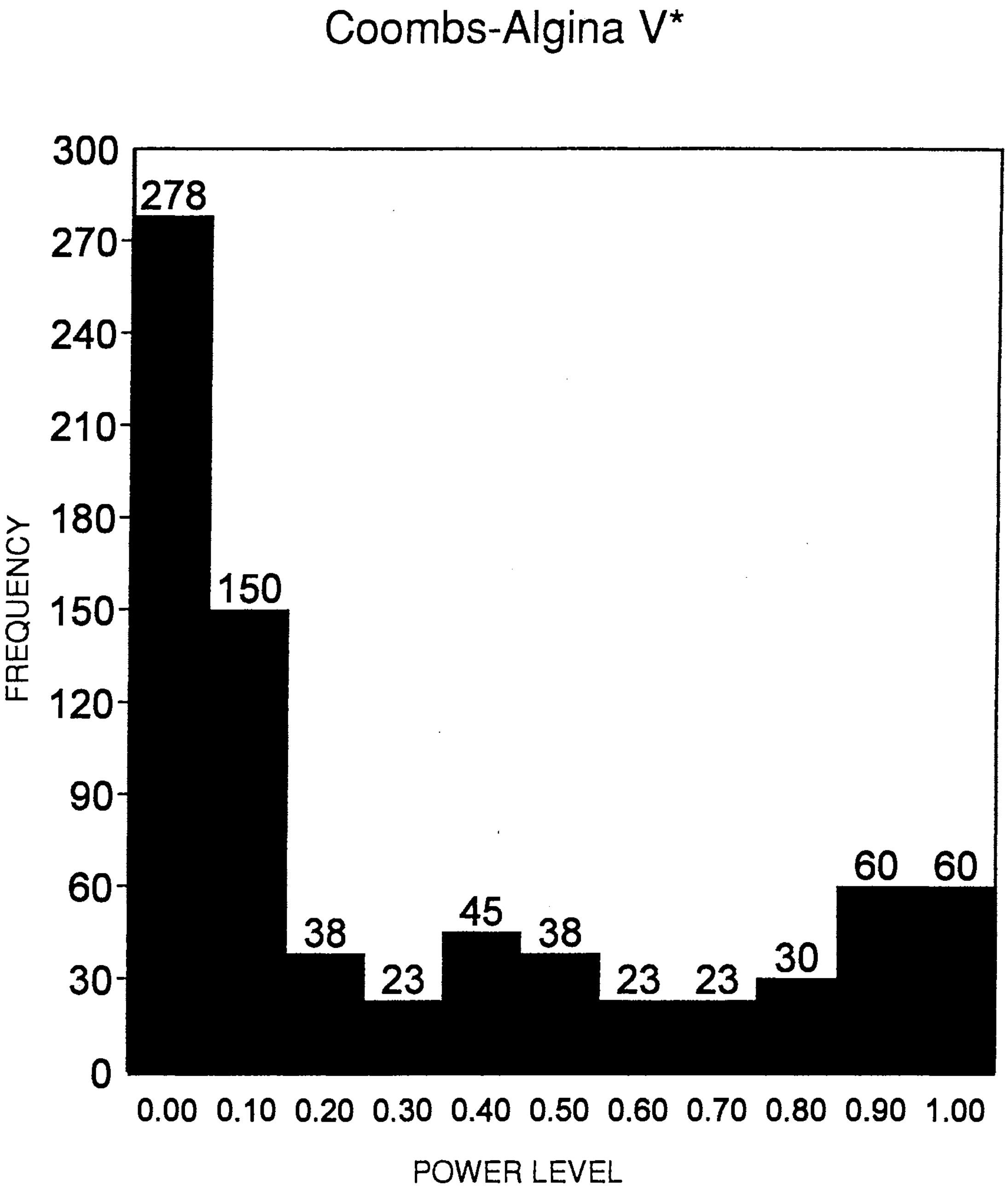


Figure 40. Frequency histogram of estimated power levels for the Coombs-Algina V* test.

the histograms and percentile table indicate that (a) power levels for all tests across all conditions tend to be inadequate and (b) power levels by themselves do not identify any test as vastly superior to the others.

The histogram results for all six tests reveal a similar pattern. All have large frequencies at or below .10 followed by an erratic decline and leveling off between .10 and .90, and finally a rather sharp increase for .90 and above. The percentile table confirms this pattern. It reveals further that if acceptable power is defined by approximately .60 and above, it is achieved only about one-quarter of the time. Acceptable power has no generally accepted cutoff point. Cohen (1992) does, however, suggest .80 and above as acceptable power, because that level yields a ratio of 4 to 1 for Type II error rate to Type I error rate. Given the low power levels of currently available criteria, doubling that ratio to 8 to 1 (resulting in a power level of .60) may be the only way for the testing of multivariate omnibus hypotheses under assumption violations to continue. The user, however, should be aware that a Type II error will occur on the average eight times as often as a Type I error. And this "acceptable" situation will be achieved only about one-quarter of the time. The performance of the Coombs-Algina $\underline{V}^*$ test is slightly worse, while the other five tests perform somewhat better than the acceptable .60 twenty-five percent of the time. In the case of the Pillai-Bartlett and Johansen tests, some of the power advantage over the Coombs-Algina tests can be explained by their inflated Type I error rates. Adjusting for differences in Type I error rate would tend to equalize power levels. So, concern should center on identifying those conditions that maximize power while maintaining adequate Type I error control, since, in at least the conditions studied, test criteria does not appear to be a distinguishing factor in determining power level.

As with Type I error rate power was further analyzed using a split-plot analysis of variance model. Eight between factors--the same seven used in analyzing Type I error rate plus type of noncentrality structure c (concentrated or diffuse)--and one within factor (test criterion) were included in the model. The highest order interaction term was used as the error term. Hence, for the between analysis the mean squared error for the effect DT × k × p × F × r × d × s × c was the error term. The mean square for the interaction of these factors and test criterion (T) served as the error term for the within analysis.

Practical significance for an effect was measured, as in the Type I error rate analysis, using $\hat{\omega}^2$, the estimated proportion of total variance accounted for by that effect. Five factors, all between effects, accounted for over 93% of the total variance. These effects and their $\hat{\omega}^2$ values appear in Table 11. Included are all effects that accounted for at least 1% of the total variance. Two effects subsumed all others that were both statistically and practically significant:

$$k \times DT$$

$$DT \times c.$$

Table 11

Proportion of Variance in Estimated Power Accounted for by Statistically and Practically Significant Effects

	Effect	Proportion of Total Variance
Between	k	.01036
	DT × k	.01415
	DT × c	.12770
	c	.12770
	DT	.65081
		.93072

The two-way interaction k × DT and the main effects k and DT that it subsumes account for 67.532% of total variance. The effect DT × c and the two main effects it subsumes, DT and c, account for 90.621%. Table 12 shows how these percentages were computed.

Table 12

Proportion of Variance in Power Levels Accounted for by Various Effects by Group

Group	Effect	Proportion of Total Variance
k × DT	k × DT	.01415
	k	.01036
	DT	.65081
		.67532
DT × c	DT × c	.12770
	c	.12770
	DT	.65081
		.90621

The results of the split-plot analysis of variance are consistent with the results obtained from examining the percentile table and histograms for power. At least under the conditions studied, test criterion is not a practically significant factor. Identification of conditions that maximize power while adequately controlling Type I error rate should be the focal point of continued analysis using the data obtained in this study. This identification was accomplished in part using mean plots for the two practically significant interaction effects that subsume all others.

Effects of Distribution Type and Noncentrality Structure (DT and c). Mean plots for the four combinations of distribution type and noncentrality structure appear in Figure 41. It shows estimated mean power to be negligible for all

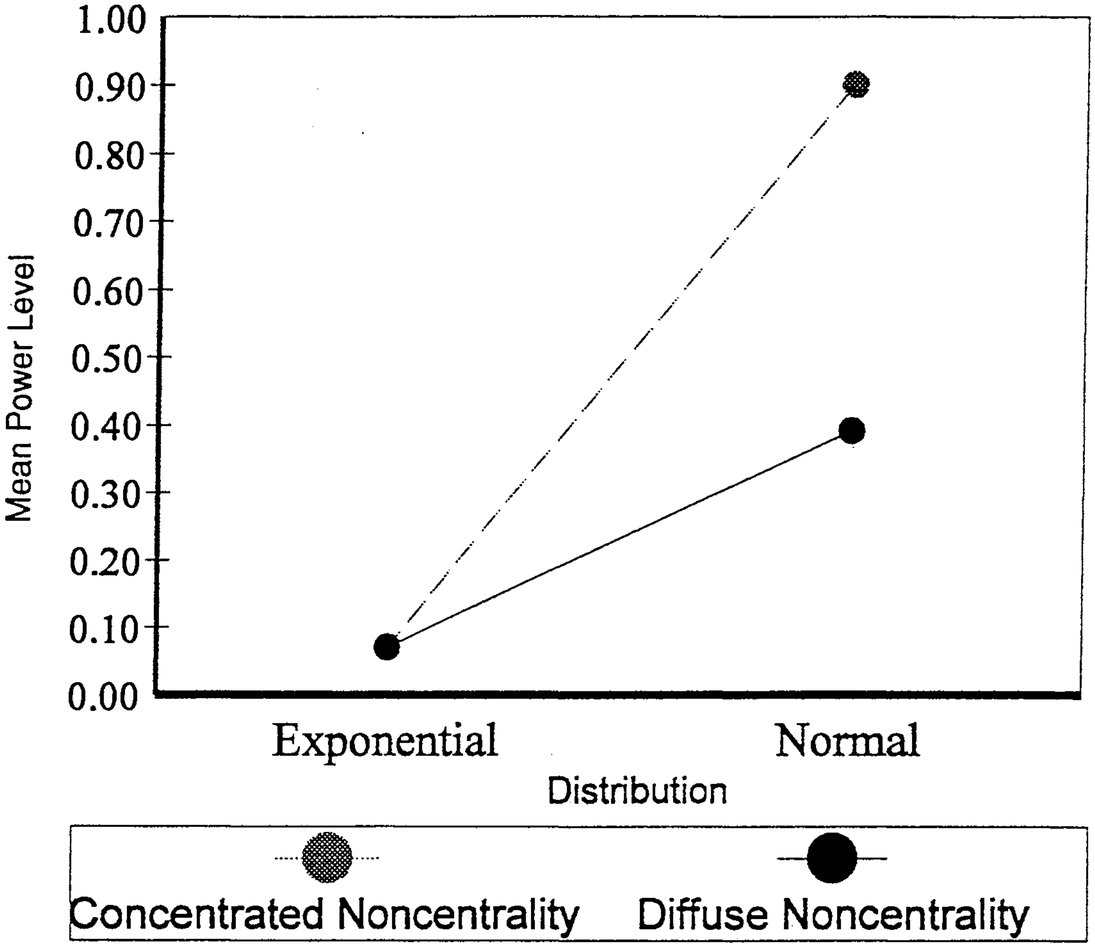


Figure 41. Estimated power levels for combinations of distributions type (DT) and noncentrality structure (c).

practical purposes when the underlying distribution is exponential, regardless of type of noncentrality structure. On the other hand, estimated power is much higher when the normality assumptions is satisfied. For the concentrated structure it is an enviable .9013. When noncentrality is diffuse estimated power falls to a usually unacceptable level of .3910. These results apply to all tests over all condition combinations, all of which include some degree of violation of the homoscedasticity assumption.

Effects of Number of Groups and Distribution Type (k and DT). The k × DT interaction was also examined using mean plots. The plots appear in Figure 42. As with the DT × c interaction, when DT = exponential, estimated mean power is woefully inadequate and precludes the use of any of the tests. When the normality assumption is met, mean power rises dramatically. When six groups are sample (k = 6) under normality, estimated mean power is .5673. For three groups estimated mean power is a perhaps acceptable .7249.

Combined Results of Type I Error Rate and Power

Because the selection of an appropriate multivariate omnibus test depends upon both Type I error rate and power level, a graphical tool that incorporates both ideas would be useful. A double box plot does so. Double box plots for the six test criteria in this study appear in Figures 43-48. In each plot 50% of all estimated Type I error rates fall within the interval delineated by the vertical sides of the box. The "whiskers" that extend left and right indicate the location of the other half of the values. Similarly, 50% of all estimated power levels are located in the interval defined by the horizontal sides of the box, the remaining 50% in the intervals described by the whiskers. Dotted lines within the box indicate medians. The optimal situation is a very small box located high on the chart (at or above the acceptable power level) and near the nominal Type I error

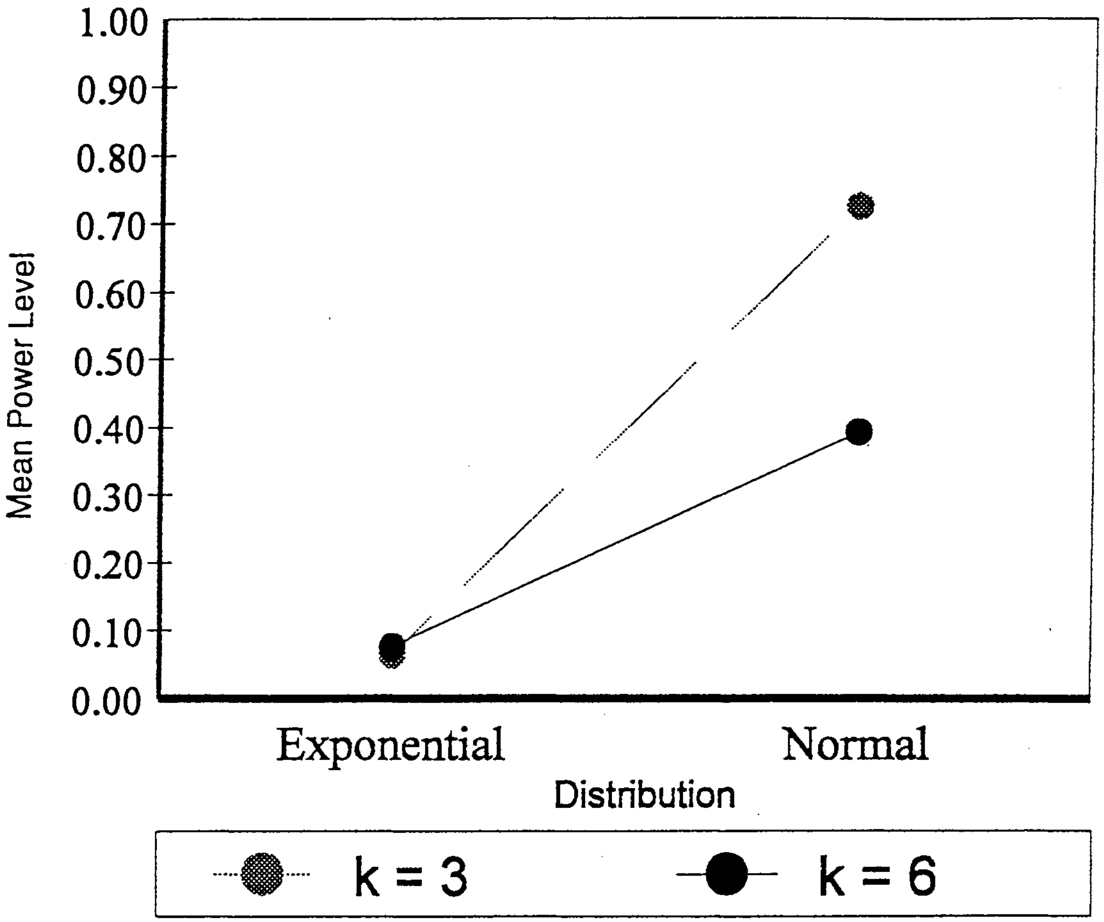


Figure 42. Estimated power levels for combinations of distributions type (DT) and number of groups (k).

rate horizontally. Further, shorter whiskers are more desirable than longer ones as they indicate a higher degree of consistency.

Figures 43-48 illustrate the superiority of the Coombs-Algina tests over both the Pillai-Bartlett and Johansen tests in controlling Type I error rate in the studied conditions. Both box widths and left-right whisker extensions confirm this result. The large heights and large up-down whisker extensions indicate the overall failure of the tests to achieve adequate power levels. Drawings of this type may be useful to future researchers as investigation continues.

Summary

The results of this study show that under the heteroscedastic experimental conditions studied (a) neither the Pillai-Bartlett test nor the Johansen test is effective in controlling Type I error rate, (b) the four Coombs-Algina tests are generally effective in controlling Type I error rate, (c) the differences among the four Coombs-Algina tests are small, and (d) none of the six tests studied offers sufficient power to advocate its use in all cases.

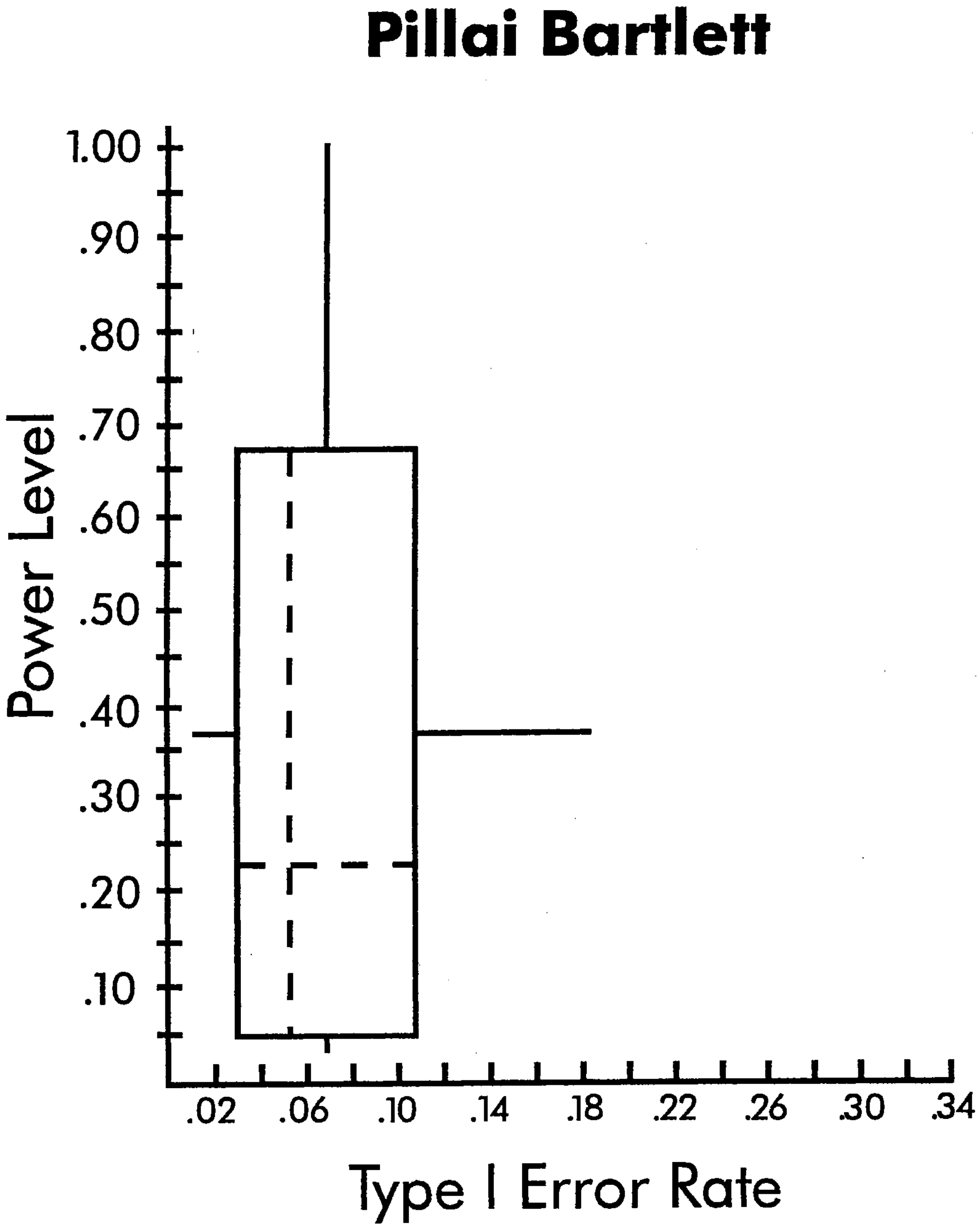


Figure 43. Double box plot of estimated Type I error rate and estimated power level for the Pillai-Bartlett test.

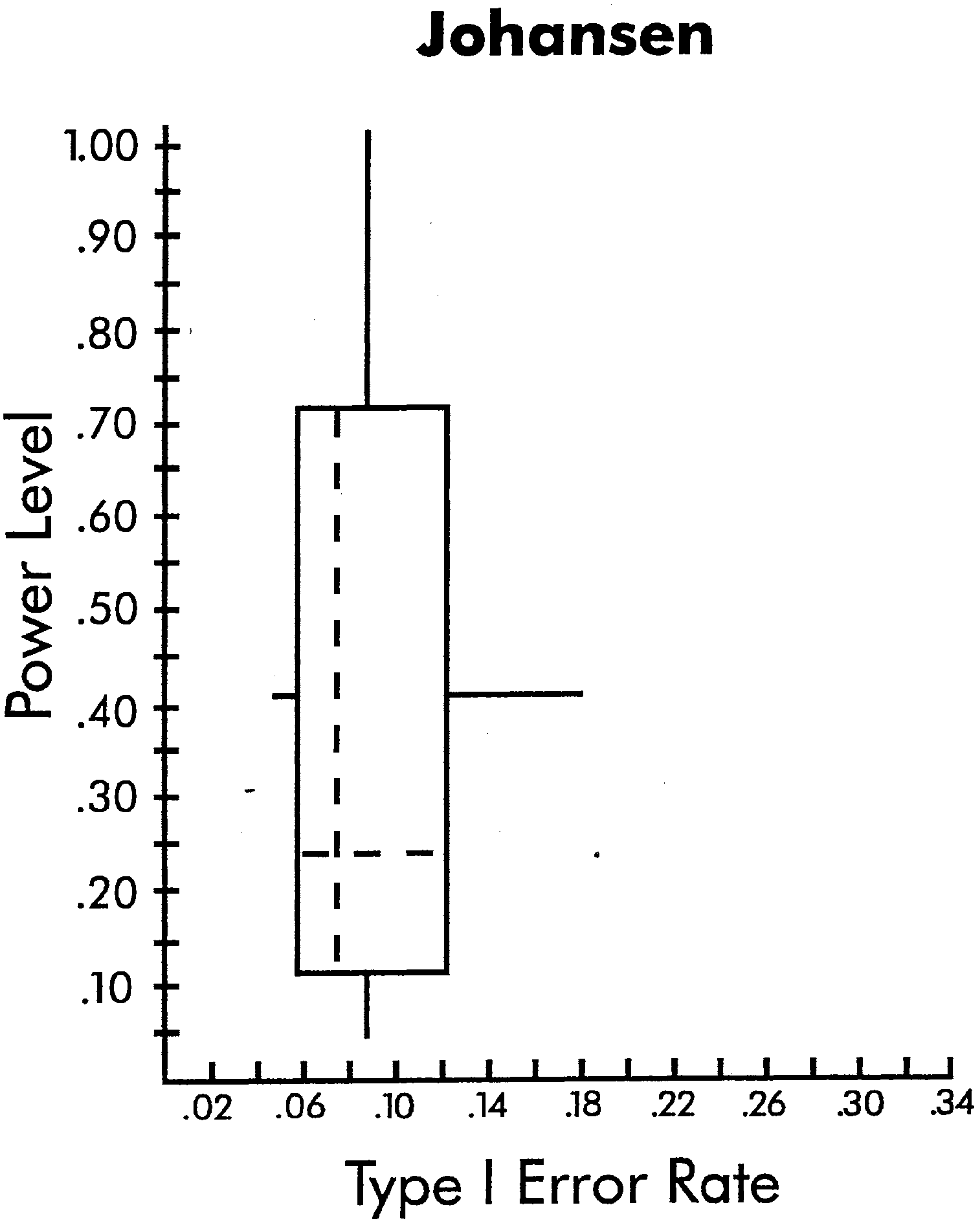


Figure 44. Double box plot of estimated Type I error rate and estimated power level for the Johansen test.

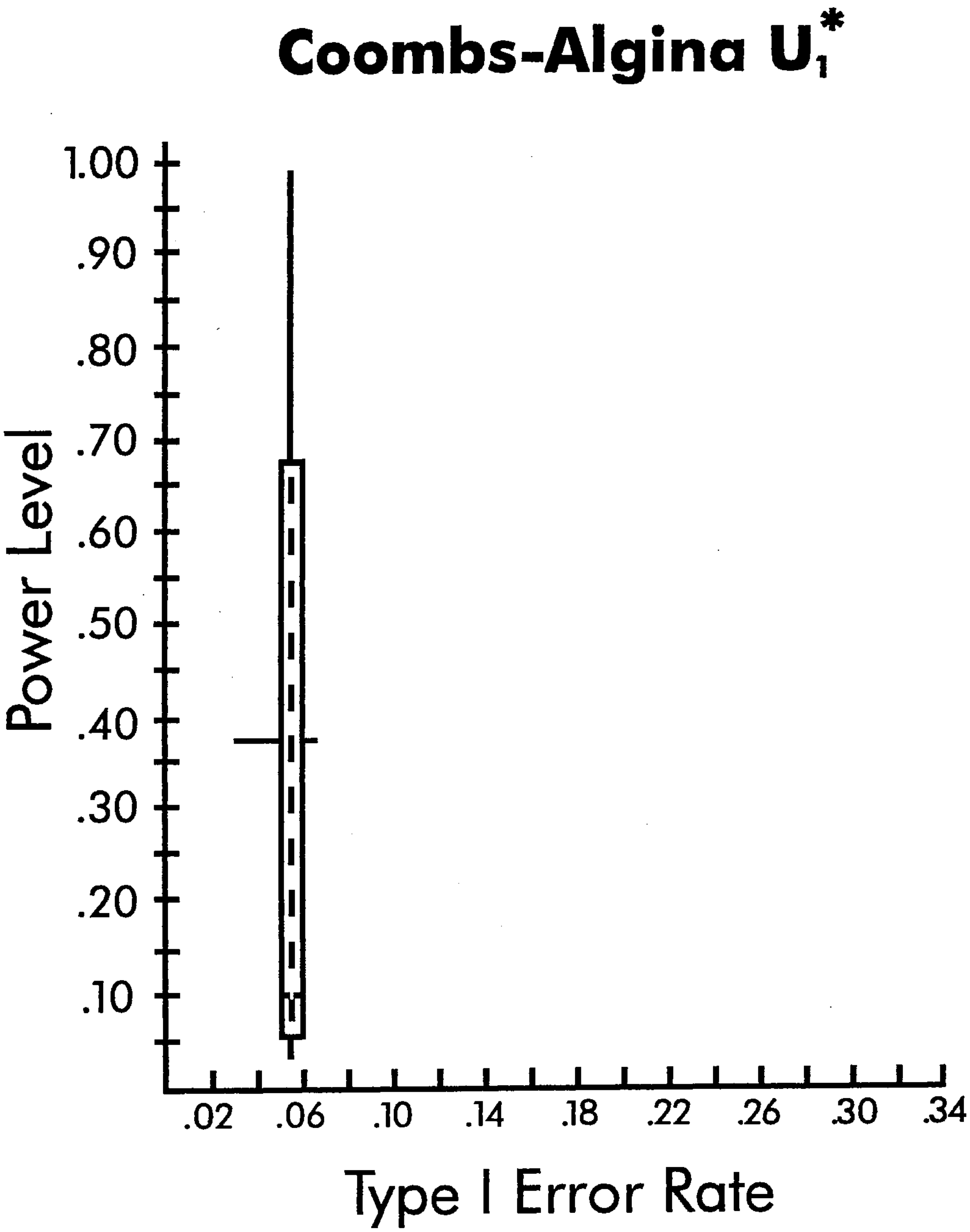


Figure 45. Double box plot of estimated Type I error rate and estimated power level for the Coombs-Algina U_1^* test.

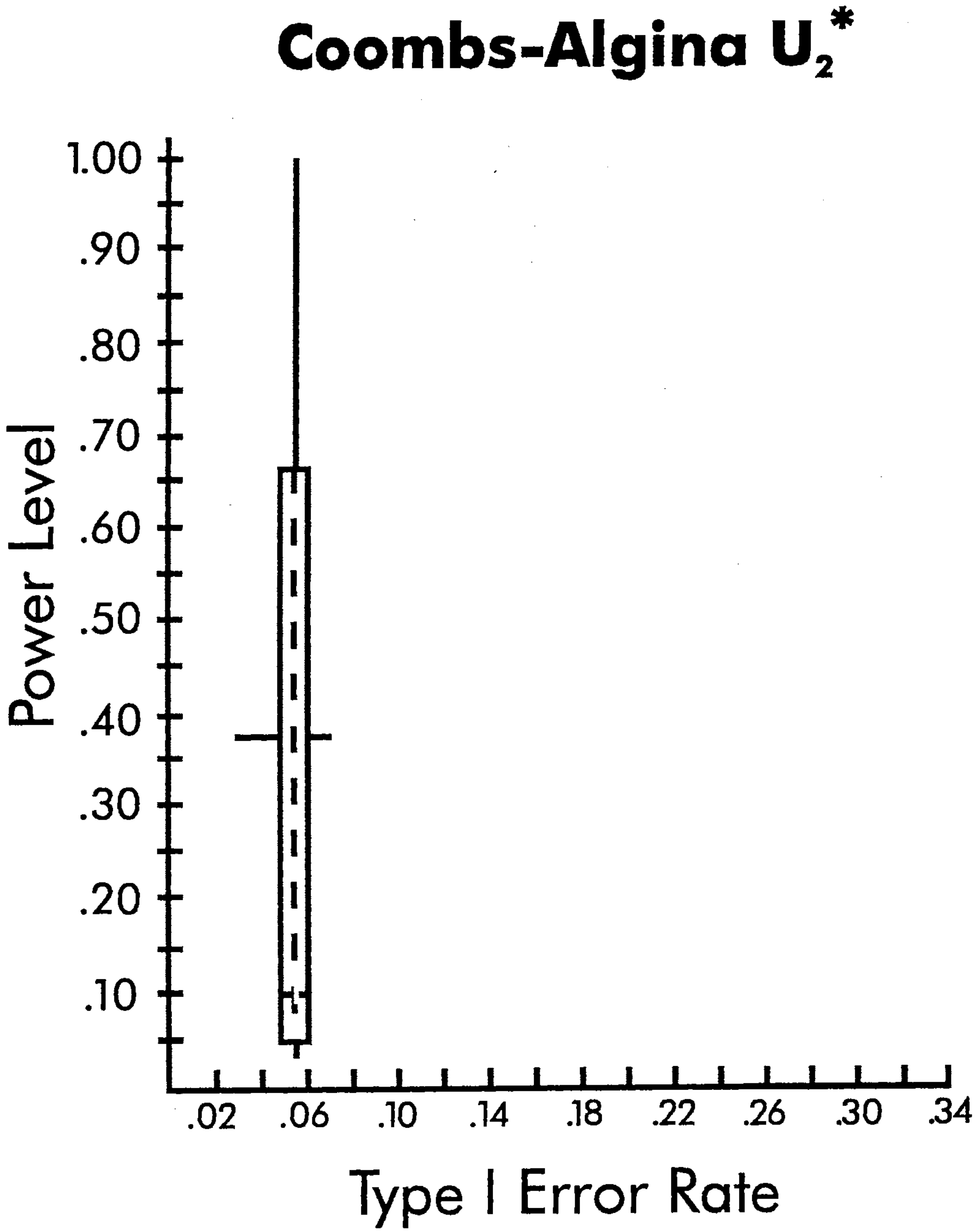


Figure 46. Double box plot of estimated Type I error rate and estimated power level for the Coombs-Algina U_2^* test.

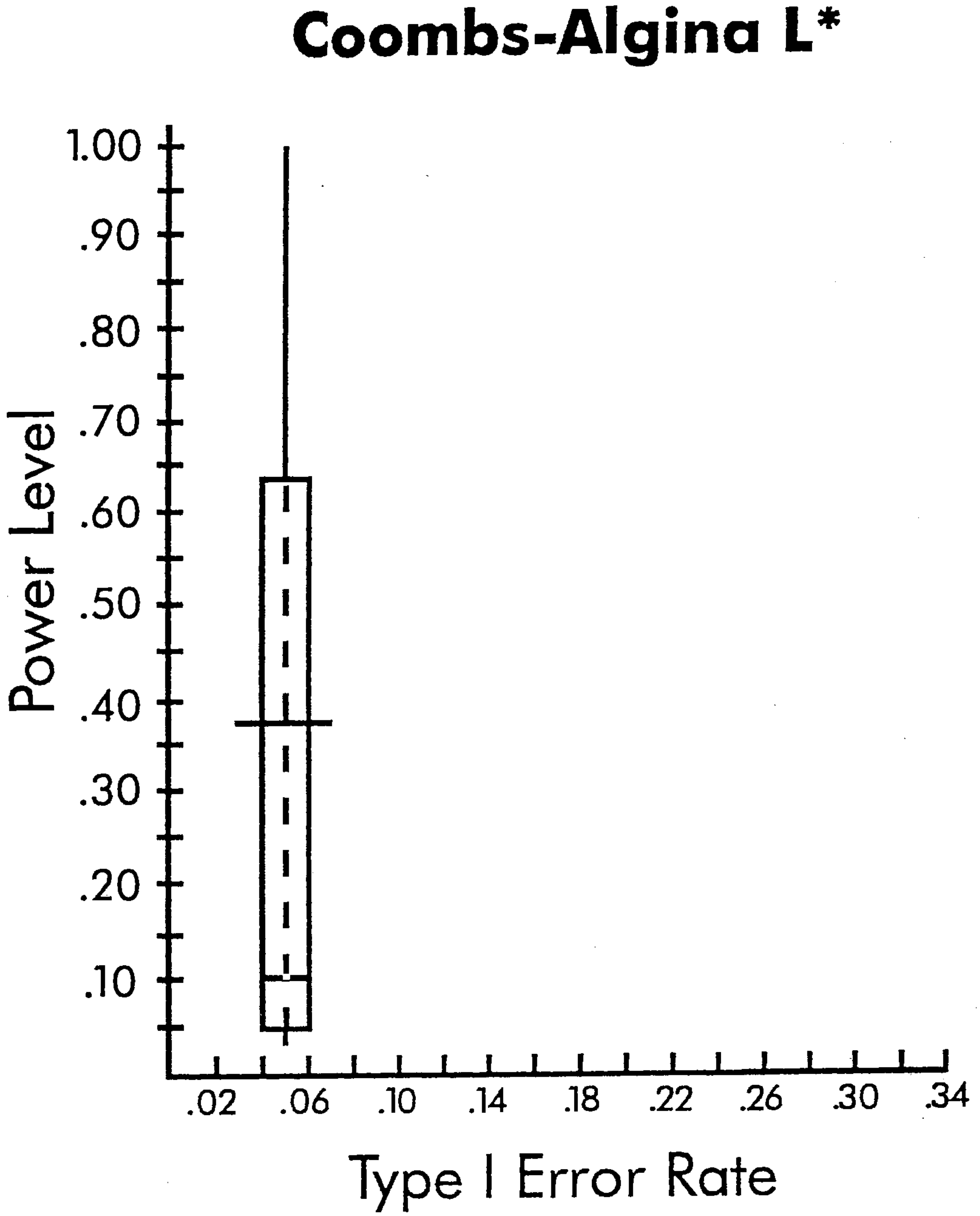


Figure 47. Double box plot of estimated Type I error rate and estimated power level for the Coombs-Algina L* test.

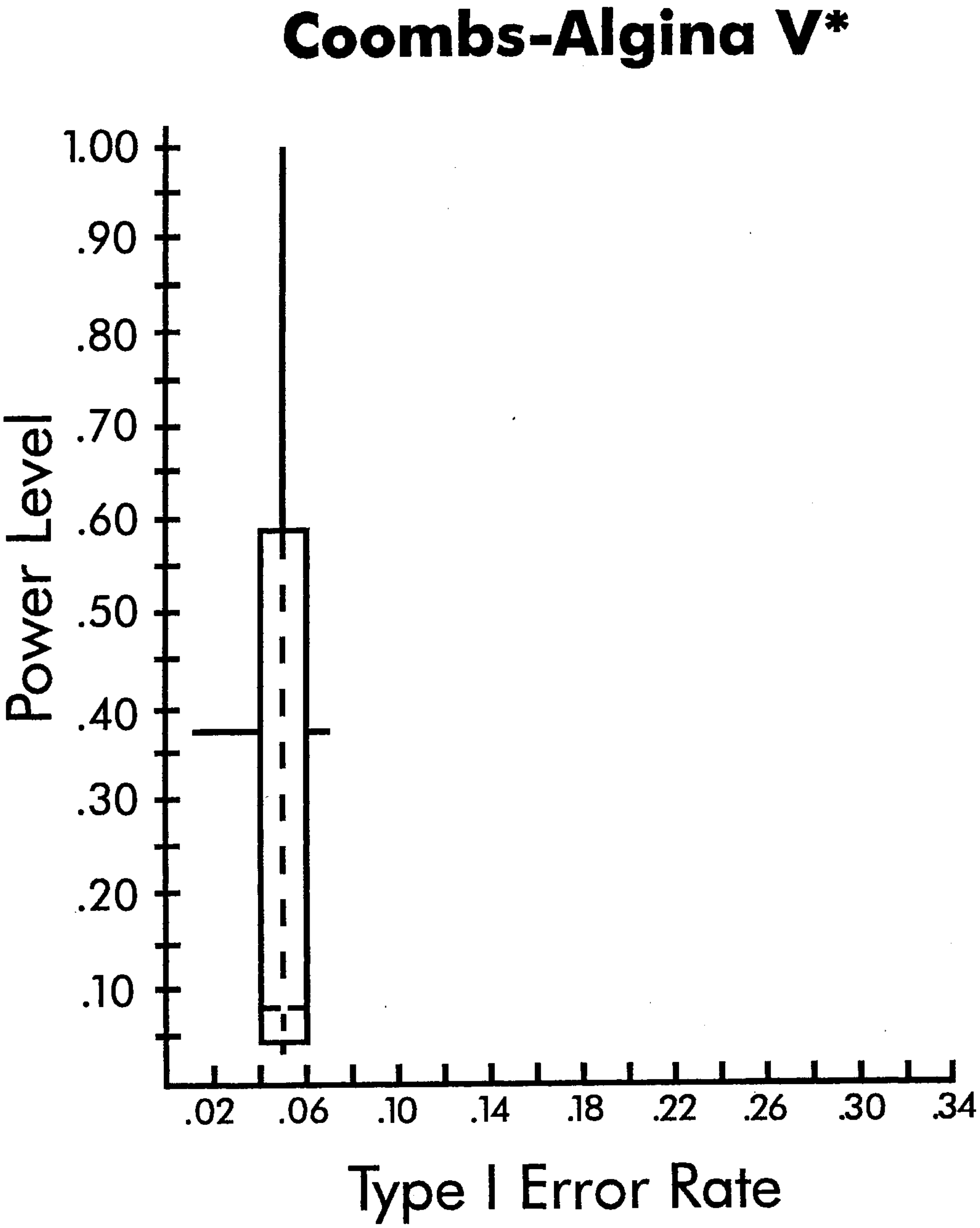


Figure 48. Double box plot of estimated Type I error rate and estimated power level for the Coombs-Algina V* test.

Chapter 5
Discussion

In partial answer to the research questions posed in Chapter 1 conclusions are offered both in terms of adequately controlling Type I error rate and maintaining suitable power levels. More complete answers to those questions can be obtained only through continued research.

Conclusions Regarding Control of Type I Error Rate

Six conclusions were drawn concerning the effectiveness of the six criteria considered in maintaining nominal Type I error rates.

Conclusion 1. The Johansen test does not provide adequate control of Type I error rates over the entire range of conditions studied. It is adequate when sampling is from a small ($\underline{k} = 3$) number of normal distributions ($\underline{DT}$ = normal) and the ratio of the smallest sample size to the number of dependent variables is large ($\underline{r} \geq 4$). Otherwise it is liberal. This conclusion is consistent with those of Coombs and Algina (in press). When the ratio of the smallest sample size to the number of dependent variables is large, the Johansen test may perform better in the positive condition.

Conclusion 2. The Pillai-Bartlett test does not provide adequate control of Type I error rates over the entire range of conditions studied. Because of the large variability it exhibits in Type I error rates when assumptions are not met – it may be very liberal or very conservative, it is not recommended when assumptions are violated.

Conclusion 3. The performances of the Coombs-Algina tests suggest adequate control of Type I error rate over the set of experimental conditions included in this study. $\underline{V}^*$ tends to be slightly conservative. $\underline{U}_1^*$ and $\underline{U}_2^*$ tend to be slightly liberal. Of the four Coombs-Algina tests $\underline{L}^*$ is the most effective overall in maintaining nominal Type I error rates in the studied conditions.

Condition 4. $\underline{V}^*$ is the best choice for controlling τ when sampling from a large number of groups. For all Coombs-Algina tests actual Type I error rate increases as the number of groups sampled increases. The increase is smallest for $\underline{V}^*$. For six populations $\underline{V}^*$ is the most effective of the Coombs-Algina tests except when the ratio of the smallest sample size to the number of dependent variables is 2. Under normality $\underline{V}^*$ is the most effective of the tests when sampling from six populations. The conservative tendency of $\underline{V}^*$ coupled with the tendency for τ to increase with the number of groups suggests $\underline{V}^*$ may work well in controlling τ with even larger numbers of populations.

Condition 5. $\underline{V}^*$ is the best choice for controlling τ when the ratio of the smallest sample size to the number of dependent variables is large. The increases in mean values of $\hat{\tau}$ decreased in this study as the ratio increased causing the mean $\hat{\tau}$s to level off. This, coupled with the conservative tendency of $\underline{V}^*$, suggests that $\underline{V}^*$ may also perform well with larger ratios.

Condition 6. Of the four Coombs-Algina tests $\underline{L}^*$ offers the best protection against the effects of high heteroscedasticity in the negative condition. In the positive condition $\underline{V}^*$ offers the best protection.

Conclusions Regarding Power

Conclusions for any test in terms of yielding sufficient power levels are meaningful only if the test adequately controls Type I error rates. Hence, the following conclusions apply to all four Coombs-Algina tests and to the Johansen test in those conditions in which it adequately controls Type I error rates.

Conclusion 1. None of the tests possesses suitable power levels for use when underlying distributions are as skewed as the exponential distribution used in this study.

Conclusion 2. The four Coombs-Algina tests and the Johansen test (when it adequately controls Type I error rates) possess suitable power levels to detect concentrated noncentrality of the type and magnitude used in this study when distributions are normal.

Conclusion 3. The four Coombs-Algina tests and the Johansen test (when it adequately controls Type I error rates) possess only marginally adequate power levels to detect diffuse noncentrality of the type and magnitude used in this study when distributions are normal.

Conclusion 4. Sampling from a small number of normal populations maximizes power.

Limitations of This Study

The results obtained and conclusions drawn may be applied only to experiments in which the conditions match or are similar to those used in this study. Generalization is limited by the range of values assigned to (a) the distribution type, (b) the number of groups sampled, (c) the number of dependent variables, (d) the form of the sample size ratio, (e) the ratio between the smallest sample size and the number of dependent variables, (f) the degree of heteroscedasticity, (g) the relationship between sample sizes and covariance matrices, and (h) the type and magnitude of the deviation from the null hypothesis (noncentrality structure).

Suggestions for Further Research

This study has both extended earlier research and established some boundaries to guide future research. Numerous avenues for continued inquiry into the performance of the Coombs-Algina tests are suggested.

First, consistently good results in both Type I error rate control and power were obtained only with sampling from normal distributions. While the Coombs-Algina tests maintained control of τ in the extremely skewed exponential distribution, the power dropoff was so dramatic as to render the tests useless. It is reasonable to conjecture that reducing skewness will have a positive effect on power. However, empirical confirmation is required along with evidence that such changes will allow Type I error control to be maintained. The amount of reduction in skew required for satisfactory power is also an issue of interest.

Second, the effects of a wider variety of covariance matrices should be investigated. The Coombs-Algina tests, especially $\underline{L}^*$ and $\underline{V}^*$, appear to offer some immunity to the effects of heteroscedasticity. At what levels and under which combinations with other factors such as distribution type, sample size ratio form, and relationship with sample size might the immunity disappear? If the immunity is preserved, the effect of the heteroscedasticity on power levels should be examined. Although the purpose of this study was to investigate test criteria performances under assumption violations, the behaviors of the Coombs-Algina tests when all assumptions are satisfied remains an unexplored area.

Third, the effects upon Type I error control and power maintenance for increased numbers of populations deserves study. Interestingly, mean estimated Type I error rate declined for all four Coombs-Algina tests when the population number increased from three to six when sampling was done from exponential populations. The more expected result of an increase in τ occurred under normality. Type I error rates, however, remained within acceptable limits for all tests even with the increase in number of groups. These tests need to be examined for larger numbers of populations to ascertain whether adequate control continues to be maintained. Given the decrease in mean τ that occurred in this study when number of groups was increased with the exponential distribution, it remains unanswered how increases in number of groups would affect τ if sampling were done from various other distributions.

Fourth, the behaviors of the Coombs-Algina tests as the ratio between the smallest sample size and the number of dependent variables increases should be pursued. The results appear to challenge the commonly held notion that increased sample size automatically reduces error rate and increases power. Unlike the Pillai-Bartlett and Johansen tests in which mean estimated Type I error rate was inversely related to the ratio of smallest sample size to number of dependent variables, mean $\hat{\tau}$ increased with the ratio in this study for the Coombs-Algina tests. That is, as more and more observations per dependent variable appeared in the smallest sample, mean estimated Type I error rate actually increased for the Coombs-Algina tests. The pattern suggests that mean $\hat{\tau}$ will level off or approach some limiting value. Further, the limiting value, if one exists, appears to differ from test to test and may be dependent upon the number of groups or the relationship between sample sizes and covariance matrices. These relationships and the ability of the Coombs-Algina tests to maintain acceptable Type I error rates and suitable power as the rate increases offer rich research opportunities.

Fifth, the powers of the Coombs-Algina tests to detect deviations from the null hypothesis in a wider variety of ways remains an open area of research. In normal distributions power was shown to be high (.9013) for the highly concentrated structure considered, but only marginal (.3910) for the diffuse structure. Numerous structures fall between these two. Olson (1974) identified a third structure, an alternate concentrated structure. This third structure and other intermediate ones should be examined to learn at which point or points power begins to suffer.

Sixth, further distinguishing factors should be sought among the four Coombs-Algina tests. Although some were suggested in the conclusions, a majority of this study's results revealed similar patterns in the tests' behaviors. Additional recommendations differentiating among the tests would increase their value to the research community.

Finally, from a more practical standpoint, the simulated results of this and similar studies could be used in other analyses such as a regression analysis for predicting both Type I error rate and power. One might envision a computer software program in which the practitioner obtains confidence intervals for both actual Type I error rate and power level under specified conditions. The user could both predict rates and levels for already designed experiments or, preferably, design experiments to control error rate and maximize power.

The tools for inquiry of the type pursued in this study have only recently become accessible to a large number of researchers. Hence, this research and its results provide only a skeleton to help direct future investigations. The opportunities are rich and varied, allowing for investigations across a wide spectrum of data types, both "real-world" and contrived and analyses of complex relationships and interactions that continually will provide better matches with real-world events and extend the frontiers of knowledge.

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. & Tang, K.L. (1991). Robustness of Yao's, James' and Johansen's tests under variance-covariance heteroscedasticity and nonnormality. Journal of Educational Statistics, 16, 125-139.

Algina, J., & Tang, K.L. (1988). Type I error rates for Yao's and James' 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: John Wiley & Sons.

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.

Aspin, A.A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. Biometrika, 36, 290-296.

Bartlett, M.S. (1935). The effect of non-normaility on the t distribution. Proceedings of the Cambridge Philosophical Society, 31, 223-231.

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

Behrens, W. V. (1929). Ein beitrag zur fehlerberechnung bie wenigen beobachtungen. Landwirtsch Jahrbucher, 68, 807-837.

Bennett, B.M. (1951). Note on a solution of the generalized Behrens-Fisher problem. Annuals of the Institute of Statistical Mathematics, 25, 87-90.

Best, D.J., & Rayner, J.C.W. (1987). Welch's approximate solution for the Behrens-Fisher problem. Technometrics, 29, 205-210

Bird, K.D. & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. Psychological Bulletin, 93, 167-179.

Blair, R. C. (1981). A reaction to "Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance." Review of Educational Research, 51, 499-507.

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

Box, G. E. P. (1954). 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. Annuals of Mathematical Statistics 25, 290-302.

Box, G. E. P., & Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society (Series B), 17, 1-34.

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

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

Budescu, D. V. (1882). The power of the F test in normal populations with heterogeneous variances. Educational and Psychological Measurement, 42, 409-416.

Budescu, D. V., & Appelbaum, M. I. (1981). Variance stabilizing transformations and the power of the F test. Journal of Educational Statistics, 6, 55-74.

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

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfield. Biometrics, 2, 22-38.

Cochran, W. G. (1954). Some methods for strengthening the common X^2 tests. Biometrics, 10, 417-451.

Cochran, W. G., & Cox, G. M. (1950). Experimental designs. New York: John Wiley & sons, Inc.

Cohen, J. (1988) Statistical power analysis for the behavioral sciences, (2nd ed.) Hillsdale, N. J.: Erlbaum.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

Coombs, W. T. (1993). Solutions to the multivariate g-sample Behrens-Fisher problem based upon generalizations of the Brown-Forsythe F* and Wilcox H_m tests. Dissertation Abstracts International, 51, 313b. (University Microfilms No.9314213).

Coombs, W. T. & Algina, J. (in press). New test statistics for MANOVA/descriptive discriminant analysis. Educational and Psychological Measurement.

Coombs, W. T., & Algina, J. (1994). Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal. Manuscript submitted for publication.

Cressie, N.A.C., & Whitford, H. J. (1986). How to use the two sample t-test. Biometrical Journal, 2, 131-148.

Daniels, H.E. (1938). The effect of departures from ideal conditions other than non-normality on the t and z tests of significance. Proceedings of the Cambridge Philosophical Society, 34, 321-328.

David, F. N., & Johnson, N. L. (1951). The effect of non-normailty on the power function of the F-test in the analysis of variance. Biometrika, 38, 43-57.

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

Donaldson, T. S. (1968). Robustness of the F-test to errors of both kinds and the correlation between the numerator and denominator of the F-ratio. American Statistical Journal, 63, 660-676.

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.

Fenstad, G. W. (1983). A comparison between the U and V tests in the Behrens-Fisher problem. Biometrika, 70, 300-302.

Fisher, R. A. (1935). The fiducial argument in statistical inference. Annals of Eugenics, 6, 392-398.

Fisher, R. A. (1939). The comparison of samples with possible unequal variances. Annals of Eugenics, 9, 174-180.

Fisher, R. A., & Healy, M. R. J. (1956). New tables of Behrens test of significance. Journal of the Royal Statistical Society (Series B), 18, 212-216.

Frieman, J. A., Chalmers, T.C., Smith, H., & Kuebler, R. R. (1978). The importance of beta, the Type II error and sample size in the design and interpretation of the randomized control trial. The New England Journal of Medicine, 299, 690-694.

Gabriel, K. R. (1968). Simultaneous test procedures in multivariate analysis of variance. Biometrika, 55, 489-504.

Games, P.A., & Lucus, P. A. (1966). Power of the analysis of variance of independent groups on non-normal and normally transformed data. Educational and Psychological Measurement, 26, 311-327.

Gayen, A. K. (1950). Significance of difference between the means of two non-normal samples. Biometrika, 37, 399-408.

Gnandesikan, R., Lauh, E., Snyder, M. & Yao, Y. (1965). Efficiency comparisons of certain multivariate analysis of variance test procedures (Abstract). Annals of Mathematical Statistics, 36, 356-357.

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

Gronow, D.G. C. (1951). Test for the significance of the difference between means in two normal populataions having unequal variances. Biometrika, 38, 252-256.

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.

Hampel, F. R., Ronchetti, E.M., Rouseeuw, P. J., & Stahel, W. A. (1986). Robust statistics: The approach based on influence functions. New York: John Wiley & Sons.

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 affects 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. American Statistical Association Journal, 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. Proceedings of the Berkeley Symposium of Mathematical Statistics and Probability, 2, 23-41.

Hsiung, T., Olejnik, S., & Huberty, C.J. (1994). Comment on a Wilcox test for comparing means when variances are unequal. Journal of Educational Statistics, 19, 111-118.

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

Hsu, P. L. (1938b). Notes on Hotelling's generalized t. Annals of Mathematical Statistics, 11, 231-243.

Hsu, P. L. (1940). On generalized analysis of variance (I). Biometrika, 31, 221-237.

Hughes, D. T., & Saw, J.G. (1972). Approximating the percentage points of Hotelling's generalized T^2 statistic. Biometrika, 59, 224-226.

Ito, K. (1960). Asymptotic formulae for the distribution of Hotelling's generalized T^2 statistic, II. Annals of Mathematical Statistics, 27, 1148-1153.

Ito, K. (1962). A comparison of the powers of two multivariate analysis of variance tests. Biometrika, 49, 455-462.

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, 51,71-82.

James, G.S. (1951). The comparison of several groups of observations when the ratios of the 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.

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

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

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

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

Lauter, J. (1978). Sample size requirements for the T^2 text of MANOVA (tables for one-way classification). Biometrika Journal, 20, 389-406.

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.

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

Lee, Y. (1972). Some results on the distribution of Wilks's likelihood-ratio criterion. Biometrika, 95, 649-664.

Levy, K. J. (1978a). Some empirical power results associated with Welch's robust analysis of variance technique. Journal of Statistical Computation and Simulation, 8, 43-48.

Levy, K. J. (1978b). An empirical comparison of the ANOVA F-test with alternatives which are more robust against heterogeneity of variance. Journal of Statistical Computation and Simulation, 8, 49-57.

Lindquist, E. F. (1953). Design and analysis of experiments in psychology and education. Boston: Houghton-Mifflin.

Marscuilo, L. A. (1971). Statistical methods for behavioral seience research. New York: McGraw Hill.

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.

McKeon, J. J. (1974). F approximations to the distribution of Hotelling's T^2. Biometrika, 61, 381-383.

McFatter, R. M., & Gollob, H. F. (1986). The power of hypothesis tests for comparisons. Educational and Psychological Measurement, 46, 883-886.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Moher, D., Dulberg, C. S., & Wells, G. A. (1994). Statistical power, sample size, and their reporting in randomized controlled trials. JAMA: Journal of the American Medical Association, 272, 122-124.

Nath, R., & Duran, B.S. (1983). A robust test in the multivariate two-sample location problem. American Journal of Mathematical and Management Sciences, 3, 225-249.

Neave, H.R., & Granger, C. W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences of mean. Technometrics, 10, 509-529.

Nel, D. G., & van der Merwe, C. A. (1986). A solution to the multivariate Behrens-Fisher problem. Communications in Statistics-Theory and Methodology, 15, 3719-3735.

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, 579-586.

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

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

Oshima, T. C., & Algina, J. (1992b). 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.

Pearson, E. S. (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. Biometrika, 21, 259-286.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. Biometrika, 23, 114-133.

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

Pillai, K. C. S. (1956). On the distribution of the largest or the smallest root of a matrix in multivariate analysis. Biometrika, 43, 122-127.

Pillai, K. C. S. (1960). Statistical tables for tests of multivariate hypotheses. Manila: Univeristy of the Phillipines, Statistical Center.

Pillai, K. C. S., & Dotson, C. O. (1969). Power comparisons of tests of two multivariate hypotheses based on individual characteristic roots. Annals of the Institute of Statistical Mathematics, 21, 49-66.

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., & Samson, P. (1959). On Hotelling's generalization of T^2. Biometrika, 46, 160-165.

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.

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

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

Pulver, A. E., Bartko, J. J., & McGrath, J. A. (1988). The power of analysis: Statistical perspectives. (Part 1). Psychiatry Review, 23, 295-299.

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.

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

Rao, C. R. (1948). Tests of significance in multivariate analysis. Biometrika, 35, 58-79.

Rao, C. R. (1952). Advanced statistical methods in biometric research. New York: Wiley

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

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? Journal of Consulting and Clinical Psychology, 58, 646-656.

Roy, S. N. (1945). The individual sampling distribution of the maximum, the minimum, and any intermediate of the p-statistics on the null hypothesis. Sankhya, 7, 133-158.

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.

Satterthwaite, F. E. (1946). An approximatae distribution of estimates of variance components. Biometrics, 2, 110-114.

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

Scheffe', H. (1943). On solutions of the Behrens-Fisher problem, based on the t-distribution. Annals of Mathematical Statistics, 14, 35-44.

Scheffe', H. (1959). The analysis of variance. Chichester: John Wiley & Sons.

Scheffe', H. (1970). Practical solutions of the Behrens-Fisher problem. Journal of the American Statistical Associatoin, 65, 1501-1508.

Srivastava, A. B. L. (1959). Effect of non-normality on the power of the analysis of variance test. Biometrika, 46, 114-122.

Stevens, J. P. (1972). Four methods of analyzing between variation for the k-group MANOVA problem. Multivariate Behavioral Research, xx, 499-522.

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

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

Stevens, J. P. (1992). Applied multivariate statistics for the social sciences, 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates.

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

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

Sukhatme, P. V. (1938). On Fisher and Behrens' test of significance for the difference in means of two normal samples. Sankhya, 4, 39-48.

Tan, W. Y. (1982). Sampling distributions and robustness of t, F, and variance-ratio in the two samples and ANOVA models with respect to departure from normality. Communications in statistics: theory and methods, 11, 2485-2511.

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

Tiku, M. L. (1971). Power function of the F-test under non-normal situations. Journal of the American Statistical Association, 66, 913-916.

Tiku, M. L. (1980). Robustness of MML estimators based on censored samples and robust test statistics. Journal of Statistical Planning and Inference, 4, 123-143.

Tiku, M. L. (1982). Testing linear contrasts of means in experimental design without assuming normaility and homogeneity of variances. Biometrical Journal, 24, 613-627.

Tiku, M. L., Gill, P. S., & Balakrishnan, N. (1989). A robust procedure for testing the equality of mean vectors of two bivariate populations with unequal covariance matrices. Communications in statistics: theory and methods, 18, 3249-3265.

Tiku, M. L., & Singh, M. (1981). Robust test s for means when population variances are unequal. Communications in statistics: theory and method, 10, 2057-2071.

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

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76, 105-110.

Wald, A. (1955). Testing the difference between the means of two normal populations with unknown standard deviations in T. W. Anderson and others, eds., Selected papers in statistics and probability by A. Wald. New York: McGraw Hill, 669-695.

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 student's' problem when several different population variances are involved. Biometrika, 34, 28-35.

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

Wilcox, R. R. (1987). New designs in analysis of variance. Annual Review of Psychology, 1987, 38, 29-60.

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

Wilcox, R. R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. Journal of Educational Statistics, 14, 269-278.

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

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

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

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, 933-943.

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.

Young, R. K., & Veldman, D. J. (1963). Heterogeneity and skewness in analysis of variance. Perceptual and Motor Skills, 16, 588.

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

Yuen, K. K., & Dixon, W. J. (1973). The approximate behaviour and performance of the two-sample trimmed t. Biometrika, 60, 369-374.

Zimmerman, D. W. & Zumbo, B. D. (1993). Rank transformations and the power of the student t test and Welch t′ test for non-normal populations with unequal variances. Canadian Journal of Experimental Psychology, 47, 523-539.