NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

A Monte Carlo Comparison of the Type I Error Rates of the Likelihood Ratio Chi-Square Test Statistic and Hotelling's Two-Sample T^2 on Testing the Differences Between Group Means

by

John R. Boulet

Faculty of Education

Thesis submitted to
the school of Graduate Studies and Research
in partial fulfilment of the requirements for the
MA degree in Education

University of Ottawa

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

ISBN 0-315-60600-2

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

ABSTRACT

The present paper demonstrates how Structural Equation Modelling (SEM) can be used to formulate a test of the difference in means between groups on a number of dependent variables. A Monte Carlo study compared the Type I error rates of the Likelihood Ratio (LR) Chi-square (χ^2) statistic (SEM test criterion) and Hotelling's two-sample T^2 statistic (MANOVA test criterion) in detecting differences in means between two independent samples. Seventy-two conditions pertaining to average sample size $((n_1+n_2)/2)$, extent of inequality of sample sizes $(n_1:n_2)$, number of variables (p), and degree of inequality of variance-covariance matrices $(\Sigma_1:\Sigma_2)$ were modelled. Empirical sampling distributions of the LR χ^2 statistic and Hotelling's T^2 statistic consisted of 2000 samples drawn from multivariate normal parent populations. The actual proportion of values that exceeded the nominal levels are presented. The results indicated that, in terms of maintaining Type I error rates that were close to the nominal levels, the LR χ^2 statistic and Hotelling's T^2 statistic were comparable when $\Sigma_1=\Sigma_2$ and $(n_1+n_2)/2$:p was relatively large (i.e., 30:1). However, when $\Sigma_1=\Sigma_2$ and $(n_1+n_2)/2$:p was small (i.e., 10:1) Hotelling's T^2 statistic was preferred. When $\Sigma_1\neq\Sigma_2$ the LR χ^2 statistic provided more appropriate Type I error rates under all of the simulated conditions. The results are related to earlier findings, and implications for the appropriate use of the SEM method of testing for group mean differences are noted.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Marc Gessaroli for his assistance throughout the study. His observations and recommendations, particularly with respect to the initial formulation of the research project, were invaluable.

The cooperation and support of Len Fleming from the Computing and Communications Services of the University of Ottawa was also greatly appreciated.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

The comparison of means of two populations on the basis of two independent samples is one of the oldest problems in statistics. Traditionally, multivariate analysis of variance (MANOVA) has been used to analyze problems which incorporate multiple outcome variables. In a one-way design this involves determining (by a statistical test) if any group differences exist on any of the dependent variables. In this type of design the hypothesis that is tested is that the populations from which the groups are selected have the same means for all dependent variables.

Recently, Structural Equation Modelling (SEM) has been applied in the analysis of experimental data from multiple groups. As a result, alternate mathematical formulations of a number of common univariate and multivariate statistical techniques, including MANOVA, have been put forth. This new methodology is convenient and flexible and can be thought of as a small part of a more general model. Furthermore, these new formulations, which need not be based on least-squares inferences, have important statistical advantages over their predecessors. Above all, they allow for the testing of mean differences in

experimental designs under a less restrictive set of assumptions.

Research on the use of SEMs in experimental designs has only begun to appear in the literature. In addition, the applicability of the technique has not been tested in many of the conditions that would be of interest to applied researchers. This is especially true with regard to the fit statistic that is commonly used to assess the adequacy of these multiple-group models. The proper use of the Likelihood Ratio Chi-square (LR χ^2) statistic depends primarily upon the assumption that the sample size is "relatively large". Although many studies have investigated the adequacy of this large sample statistic in small samples (e.g., Bearden, Sharma & Teel, 1982; Geweke & Singleton, 1980; La Du & Tanaka, 1989; Mulaik, James, Van Alstine, Bennett, Lind, & Stilwell, 1989), this research has not extended to models used to analyze experimental designs.

Multivariate Analysis of Variance and multiple-group structural equation modelling can both be used to test for group mean differences on a number of dependent variables. Both of these techniques rely, to some extent, on meeting the assumptions of their respective estimation procedures. The present study used Monte Carlo methods to study the comparative performance of these two techniques in detecting differences in means between two groups over a number of unique conditions. Specifically, the Type I error performance of Hotelling's T^2

statistic (MANOVA test criterion) and the LR χ^2 statistic (SEM test criterion) were examined under varying sample sizes, between-group variance-covariance heterogeneity and ratios of average sample size to number of dependent variables. This allowed for a quantification of the magnitude of effect of these hypothetical conditions across the two statistics.

The formulation and utilization of Hotelling's T^2 statistic is discussed in Chapter 2. This includes a review of some the factors that may influence the robustness of this statistical test. The SEM that can be used test differences in group means is also introduced in Chapter 2. This is augmented with a review of the literature regarding the use of the LR χ^2 statistic for testing the goodness-of-fit (GOF) of SEMs. A review of the specific research problem that is addressed in the present study is advanced in Chapter 3. The discussion of the methodology, focusing on the Monte Carlo design and data generation procedure, is presented in Chapter 4. A summary and interpretation of the results of the analysis is put forward in Chapter 5. The remaining 2 chapters include a discussion of the comparative utility of the SEM approach in analyzing problems that require the testing of group mean differences.

CHAPTER 2

Test Statistics

A description of the formulation and use of the two test statistics that were investigated in this study (Hotelling's two-sample T^2 statistic and the LR χ^2 statistic) is presented in this section. The relevant literature related to the behaviour of these statistics under imperfect conditions is also discussed.

Hotelling's T^2 Statistic

For the two-group p-variate design Hotelling's (1931) two-sample T^2 procedure can be used to determine the existence of significant treatment effects on the p dependent variables. It is the multivariate analog of the familiar t-ratio for testing the significance of the difference of two means that are based on two independent samples. The mathematical formulation for the two-sample T^2 test is given below:

$$T^2 = (N_1N_2 / N_1+N_2) (X_1 - X_2)'S^{-1}(X_1 - X_2), \quad (1)$$

where S is the pooled within group covariance matrix

and N_1, N_2 are the number of cases in the two samples.

Hotelling's T^2 statistic has been shown to be the uniformly most powerful test for the two-group case (Anderson, 1958, pp. 115-118).

The are a number of theoretical issues that must be addressed when using Hotelling's T^2 statistic to test for group mean differences. First, the significance test for T^2, as well as other MANOVA test criteria, is based on the multivariate normal distribution. This implies that the observations on the p dependent variables follow a multivariate normal distribution in each group. Second, it is assumed that the population variance-covariance matrices for the p dependent variables in each group are equal. Finally, it is assumed that the observations in each group are independent.

Assumption Violation

A number of theoretical and empirical studies have been undertaken in order to investigate the behaviour of the T^2 statistic under violation of the

aforementioned assumptions. A brief review of this research is now presented.

Various researchers have investigated the effect of non-multivariate normality on the Type I error rates and power of Hotelling's two-sample T^2 statistic. Ito (1969) found that for sufficiently large samples T^2 was quite robust under the violation of the normality assumption. However, since this study was theoretical in nature, no exact specification could be made as to what constituted "sufficiently large". Hopkins and Clay (1963) found, through Monte Carlo methods, that for $n_1, n_2 \geq 10$ and $p=2$ the bivariate distribution of Hotelling's T^2 statistic was not substantially affected by moderate degrees of kurtosis. These studies, combined with previous research on the topic (see Ito, 1980), suggest that, given sufficiently large sample sizes, deviations from multivariate normality should have little effect on the robustness of Hotelling's two-sample T^2 statistic.

There have also been a number of studies that have explored the effect of variance-covariance heterogeneity on the robustness of T^2. Ito and Schull (1964) investigated the large sample properties of the distribution of T^2 and found that, for sufficiently large samples, the test was robust under violation of equality of population variance-covariance matrices. Hopkins and Clay (1963) simulated the effects of inequality of variance-covariance matrices and of kurtosis on the central distribution of T^2 for $p=2$, 5, 10, and 20. In their study they drew 1000 random

pairs of samples of size n_1, n_2 from a bivariate normal population having $\sigma_1/\sigma_2=1$, 1.6 and 3.2. They found that T^2 was robust against variance-covariance heterogeneity for samples with n_1, $n_2 \geq 10$ but that this result did not extend to cases with unequal sample size. The authors concluded that the effects of inequality of variance-covariance matrices are small if $n_1=n_2$ but may be serious if the sample sizes are markedly different.

Holloway and Dunn (1967) also used Monte Carlo methods to study the effect of inequality of variance-covariance matrices on the distribution of the T^2 statistic. They drew 5,000 or 10,000 pairs of samples of size n_1, n_2 from multivariate normal distributions for p=1,2,3,5,7,10. The results indicated that the actual Type I error rate (τ) departed from the nominal rate (α) as variance-covariance heterogeneity increased. The magnitude of this departure escalated with an increase in the number of dependent variables (p). Also, if sample sizes were large, α was not seriously affected by the violation of the assumption of equality of variance-covariance matrices. This result did not extend to small samples (i.e., $n_1+n_2=20$), even when $n_1=n_2$. Similar to Hopkins and Clay (1963), it was also found that equal sample sizes helped in keeping the level of significance close to the nominal level.

Hakstian, Roed, and Lind (1979) studied the robustness of Hotelling's two-sample T^2 test with respect to the violation of the assumption of homogeneity of covariance matrices. In their investigation the authors also examined the effects of sample size, extent of inequality of sample sizes, and the number of variables on the test statistic. Empirical sampling distributions of the T^2 statistic were obtained from a large number of data sets, covering 108 separate conditions. These data sets each consisted of 2000 samples drawn from multivariate normal populations. It was found that the robustness of T^2 was not guaranteed with unequal sample sizes in the presence of variance-covariance heterogeneity. Furthermore, as the number of dependent variables increased and the discrepancy between sample sizes increased, the distortion in α levels escalated. For equal sample sizes, however, T^2 was generally robust to the homogeneity assumption, even when the ratio of sample size to the number of dependent variables was small. In particular, T^2 was able to withstand population scale differences of 1.5 on all variables when $n_1 = n_2$.

The literature concerning the sensitivity of the T^2 statistic to departures from the assumptions suggests that the test is robust when samples are large and equal. In addition, T^2 is robust with respect to the homogeneity assumption for equal sample sizes, even when the ratio of sample size to the number of dependent variables is quite small. However, when $n_1 \neq n_2$ and $\Sigma_1 \neq \Sigma_2$ Hotelling's T^2

is not robust, even for relatively mild departures. For conditions in which the sample sizes are not equal, the effect of heteroscedasticity depends on the relationship between n_1 and n_2, the number of cases in n_1 and n_2, and the magnitude of disparity between the group covariance matrices.

Alternative Tests Under Variance-Covariance Heterogeneity

There are many settings in which population variance-covariance heterogeneity would be expected. For example, in testing for achievement the homogeneity assumption would often be unrealistic when applied to students with various educational backgrounds (Muthén, 1989). The realistic possibility of population heterogeneity, combined with the fact that T^2 has been shown to be sensitive to violations of this assumption under certain conditions, has led to the development of modified test procedures. These include James' first order asymptotic series test (James, 1954) and Yao's approximate degrees of freedom test (Yao, 1965). Both of these tests attempt to modify the T^2 statistic so as to minimize the discrepancy between the actual significance level and the conditional prior probability of making a Type I error when heteroscedasticity is present. There are also a number of tests based on Scheffé's (1943) technique of randomization. These Scheffé based tests will not be reviewed in that they require randomized pairings of observations in different groups and the

elimination of data when $n_1 \neq n_2$.

Algina and Tang (1988) used Monte Carlo methods to compare the Type I error rates for Yao's and James' tests for various combinations of numbers of variables, sample size ratio (n_1:n_2), and degree of variance-covariance matrix heteroscedasticity. They found that when $n_1 = n_2$ both Yao's and James' tests were robust, except when the disparity in covariance matrices was large, there was a large number of dependent variables (p), and the ratio of average sample size to the number of dependent variables ($(n_1+n_2)/2/p$) was small. They also found that Yao's test was superior in performance to James' test and that the deviation of τ from α was larger for T^2 than for either Yao's or James' tests. The authors concluded that Yao's test can be safely used provided $p \leq 10$, $(n_1+n_2)/p \geq 10$, and the ratio of the larger to smaller sample size was 2:1 or less. Furthermore, if $(n_1+n_2)/p \geq 20$ Yao's test could be safely used provided $p=2$ and the sample size ratio is 5:1 or smaller, $p=6$ and the ratio is 3:1 or smaller, or $p=10$ and the ratio is 4:1 or smaller.

Structural Equation Modelling (SEM)

Structural equation modelling has been shown to be a powerful tool for specifying, estimating and testing many types of models. SEM can be used to

estimate a specific set of relationships among a set of measured variables and a set of latent variables. There are several variations of the general structural equation model, each with its own specific formulation and estimation procedures (e.g., Bentler, 1982, 1983a; Bentler & Weeks, 1980; Jöreskog & Sörbom, 1986; McDonald, 1978; Muthén, 1984). LISREL (Jöreskog & Sörbom, 1986) is one of a number of programs that can be used to specify, estimate, and test structural relationships. It is widely available and can be used to formulate a sizable number of models needed for applied research. Since the LISREL program is used in the present study, the later description of the SEM model used to test for group mean differences will utilize LISREL terminology.

Multiple-Group Structural Equation Modelling With Means

Although structural equation models have been used extensively in the social and behaviourial sciences, the adaptation of simultaneous equation systems to models including structured means has not been widespread. The linear structural relations methodology does, however, provide interesting possibilities for testing the equality of means and covariances in a multiple-group analysis. The parameters can, under conditions of normality and relatively large samples, be estimated with Maximum Likelihood (ML) or Generalized Least-Squares (GLS) procedures. A Likelihood Ratio (LR) χ^2 goodness-of-fit statistic can then

be used to test the null hypothesis that all the groups have the same population means on all the dependent variables.

The use of SEMs to test mean structures has generally focused on non-experimental designs. There have, however, been a number of researchers who have discussed the applicability of causal models in studies of multiple populations and experimental data (Bagozzi, 1977; Bentler, 1980, 1983b; Bentler & Weeks, 1979; Kenny, 1979; Lomax, 1983; Rock, Werts, & Flaugher, 1978; Sörbom, 1981). Alwin and Tessler (1974, 1985) illustrated how the parameters of causal models of experimentally derived data can be estimated using Jöreskog's model of covariance structure analysis. Bagozzi, Fornell, and Larcker (1981) showed how canonical correlation is a special case of a linear structural equation model. Rock et al. (1978) discussed the role of SEM in clarifying univariate and multivariate analysis of variance. Similarly, Bray and Maxwell (1982) provided a review of the path analytic representations of a number of common multivariate techniques including MANOVA, discriminant analysis, and stepdown analysis. The authors also supplied rudimentary causal diagrams for these techniques.

In order to estimate means via structural equation modelling one must make certain modifications in the general SEM model (see Jöreskog & Sörbom, 1986). These modifications yield models that are generally referred to as moment

structure models. This is due to the fact a moment matrix is used in the estimation process instead of the usual variance-covariance matrix. The matrix of first, second, and joint moments (M) is a matrix containing information on the expected values (means) of all variables as well as on their variances and covariances. A matrix of joint and second moments of the included variables can be calculated by adding the product of the variable's means to each covariance and the square of the variables mean to each variance. The matrix that is analyzed in LISREL is referred to as the augmented moment matrix (AM) and includes the variable means in the last row and column of the matrix.

$$AM = \begin{bmatrix} S+zz' & z \\ z' & 1 \end{bmatrix}, \tag{2}$$

where z is a vector of observed indicator means.

Specification for One-Way MANOVA

There are a number of possible SEM programs and specifications that can formulate one-way multivariate designs (e.g., Jöreskog & Sörbom, 1986; Muthén, 1987). A brief description of 2 possible formulations for comparing the means of dependent variables across groups will be advanced. An emphasis will be placed

on the model that is based on the direct comparison of means. This model will later be used in the analysis of simulated data.

Dummy Variable Regression Approach. A test of the equality of group means in a one-way design may be performed within the framework of ordinary least-squares estimation with dummy variables (e.g., Bagozzi & Yi, 1989; Kühnel, 1988). This latent variable formulation based on multivariate regression does, however, still require the same assumptions as traditional MANOVA (i.e., linearity, multinormality, variance-covariance homogeneity). Furthermore, all statistical inferences are based on the asymptotic properties of the LR χ^2 test. Hence, the results may only be valid with moderate to large samples. Although this type of SEM approach provides an interesting way to test for group mean differences, it yields little or no statistical advantage over the traditional MANOVA methodology.

Direct Comparison of Means. One-way MANOVA can also be formulated as a direct comparison of the means of the dependent variables between groups. Although a number of LISREL formulations can be derived to test the equality of mean structures between groups, the model proposed by Kühnel (1988) is used in the present study. This model is specified with observed variables only. The means are specified as regression coefficients (Λ's) on a latent variable (ξ) with

constant value 'one'. The ξ variable is defined as "1" through the use of a pseudovariable in the moment matrix. With this model it is possible to express multivariate analysis with latent variables based on linear representations that are no more complex than traditional MANOVA.

Consider the general model describing the scores of subject i within group j (j=1,...,g) on variable k (k=1,...,q) as

$$x_{ik}^{j} = \mu_k + \alpha_k^{j} + \delta_{ik}^{j} . \tag{3}$$

If we let

$$\lambda_k^{j} = \mu_k + \alpha_k^{j} \tag{4}$$

then

$$x_{ik}^{j} = \lambda_k^{j} + \delta_{ik}^{j} \tag{5}$$

In this model each observed value x_{ik} is decomposed as the mean λ^{j}_{k} of population j on dependent variable k and the deviation of the ith individual value from this mean. This is analogous to the decomposition in classical test theory of a measurement into "true score" and "error" components.

One can estimate (5) with LISREL by using the general measurement model

$$x^j_{ik} = \lambda^j_k \xi + \delta^j_{ik} \quad , \tag{6}$$

where x^j_{ik}, λ^j_k, and δ_{ik} are defined as before, and ξ is a column vector containing n independent latent variables.

Clearly, (6) is equal to (5) if $\xi = 1$. This is easily accomplished by introducing a new variable for each subject ($x^j_{i(q+1)}$), which is the constant, 1. Thus, if there are q dependent variables and the constant, 1, then there are q+1 equations corresponding to the measurement model (6). These are

$$x^j_{ik} = \lambda^j_k \xi + \delta^j_{ik} \tag{7}$$

(k = 1,...,q) and

$$x^j_{i(q+1)} = \lambda^j_{(q+1)} \xi + \delta^j_{i(q+1)} \tag{8}$$

where $x^j_{i(q+1)} = 1$. Now if we set $\lambda^j_{(q+1)} = 1$ and $\delta^j_{i(q+1)} = 0$, then (8) becomes

$$\xi = 1 \, . \tag{9}$$

After making the substitution (9) in (7), (7) can be expanded to become

$$
\begin{aligned}
x_{i1}^{j} &= \lambda_{1}^{j} + \delta_{i1}^{j} \\
x_{i2}^{j} &= \lambda_{2}^{j} + \delta_{i2}^{j} \\
&\vdots \\
x_{iq}^{j} &= \lambda_{q}^{j} + \delta_{iq}^{j} \quad .
\end{aligned}
$$

In the multiple-group case, a simultaneous test of means across the j groups is a test of the multivariate null hypothesis, H_o:

$$
\begin{aligned}
\lambda_{1}^{1} &= \lambda_{1}^{2} = \ldots = \lambda_{1}^{g} \\
\lambda_{2}^{1} &= \lambda_{2}^{2} = \ldots = \lambda_{2}^{g} \\
&\vdots \\
\lambda_{q}^{1} &= \lambda_{q}^{2} = \ldots = \lambda_{q}^{g} \quad .
\end{aligned}
$$

The omnibus test of the hypothesis of equal mean vectors across groups is modelled by constraining the corresponding lambda (λ) coefficients for the separate groups to be equal. If the extra restrictions are consistent with the pattern of moments in the data for each group then the differences between the actual first-order, second-order, and joint moments among the variables and the fitted moments estimated by the model will be minimal. In the common practice of SEM, a null hypothesis, H_o, is specified and tested against a general unrestricted alternative, H_a. The resulting χ^2 goodness-of-fit statistic yields the likelihood ratio test of the null hypothesis that all the groups have the same

population means on all the dependent variables.

Estimation Procedure

A moment structure model can be interpreted as a hypothesis regarding a specific set of relationships among a set of measured variables and a set of latent variables. The calculation of the unconstrained parameters in a moment structure model can be accomplished through a number of statistical estimators, including Maximum Likelihood estimation (ML). The theory for estimating and testing models incorporating means across different groups using ML has been well documented (e.g., Sörbom, 1974, 1981). Only a brief summary of the estimation procedures will be given here.

Suppose that N_j subjects ($j=1,...,g$) in g groups are measured on q variables. If the data are normally distributed and the sample sizes are large, then one can estimate the parameters of the model for each group using ML procedures by minimizing the following function:

$$F_j = \ln|\Omega_j| + tr(M_j\Omega_j^{-1}) - \log|M_j| - (q+1), \quad (10)$$

where Ω_j = the hypothesized moment matrix for population j,

M_j = sample moment matrix for group j,

q = the number of dependent variables.

The fit of a multiple-group model is calculated as the weighted average of the fit attained for each of the groups individually.

$$F = \Sigma\,[(N_j-1)/N)]F_j \quad (11)$$

The parameter estimates that are obtained are neither averages across groups nor based on separate calculations for each group. The estimation procedure simultaneously calculates values for the estimates that minimize the fitting function (F), given the free and constrained elements in the model. It should also be noted that other estimation procedures such as Generalized Least-Squares (GLS) or Unweighted Least-Squares (ULS) may also be employed. The respective functions, F, to be minimized can be found in Jöreskog & Sörbom (1986).

Likelihood Ratio χ^2 Test

One of the advantages of using ML (or GLS) estimation in the SEM approach of fitting the model to the data is that a Likelihood Ratio (LR) Chi-square (χ^2) test of the specified model is available. For example, in the two-group one-way model the closeness of the match between the 2 M's and the 2 Ω's can be assessed with a Likelihood Ratio Chi-square test once the fit is maximized. A multiple-group χ^2 can be obtained by multiplying the minimum of the fit function by the overall number of cases.

$$\chi^2 = (N)F \tag{12}$$

For large samples, the test evaluates whether the discrepancies between the moments implied by the model and the empirical moments are reasonably attributable to sampling variation.

For the g-group comparison of means the degrees of freedom for the χ^2 test is the total number of known input information (first, second, and joint-moments) minus the number of unknown parameters being estimated (t). The general formula for g groups is given below:

$$df = (g)\left\{ [\ (½)(q+1)(q+2)\ ] - 1 \right\} - t \qquad (13)$$

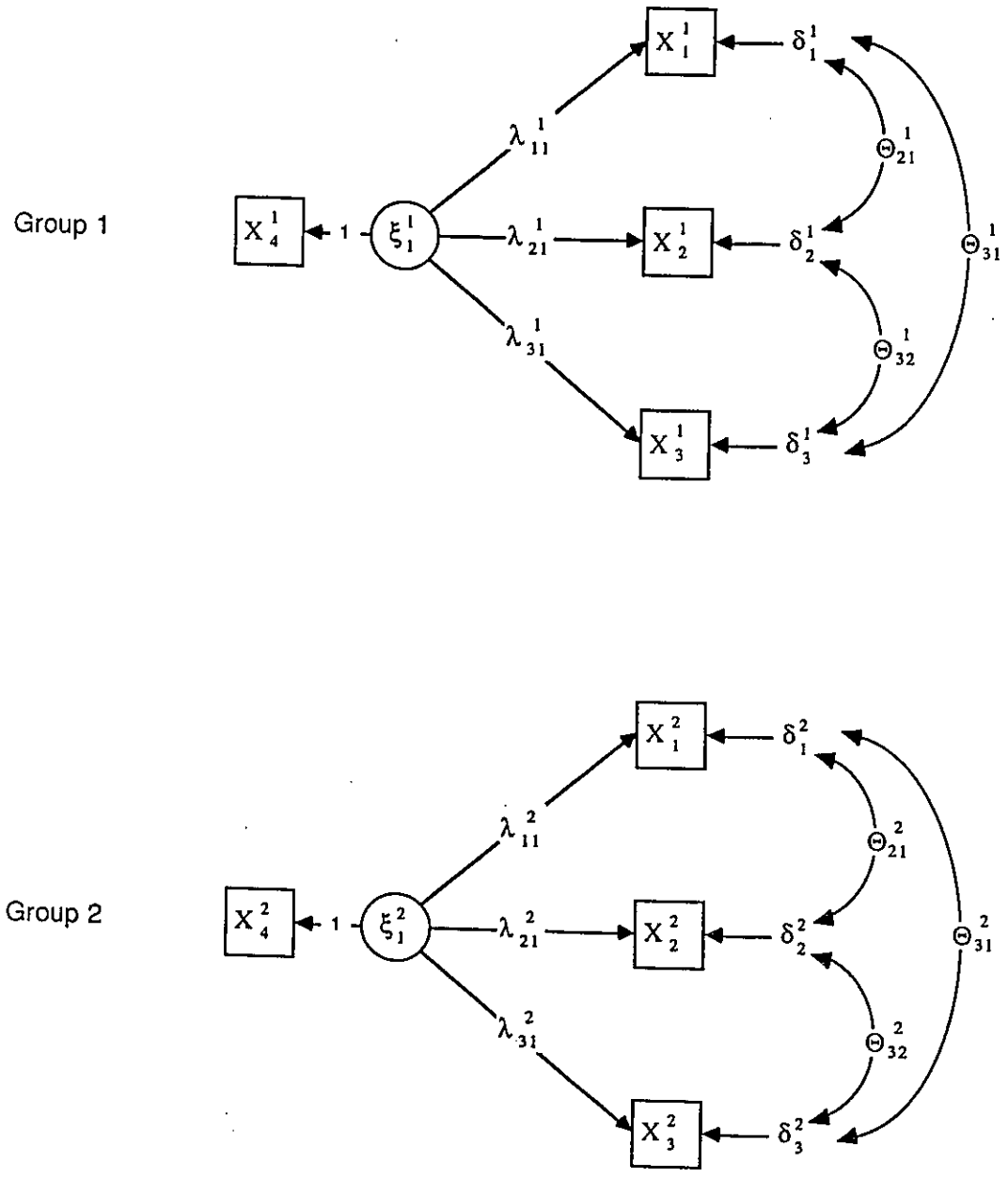
where q is the number of dependent variables in the design, and g represents the number of groups.

For example, in a two-group MANOVA design with 3 dependent variables there would be a total of 18 unique entries in the moment matrices (6 first-order moments, 6 second-order moments, and 6 joint-moments) and 15 parameters that would be estimated (3 first-order moments which are invariant over groups, 6 second-order moments, and 6 joint-moments). Therefore the degrees of freedom for the χ^2 test would be 3 (18 - 15).

For illustrative purpose the path analytic representation one-way MANOVA design with two groups is displayed in Figure 1. This example incorporates 4 indicators of the independent latent variable ξ (3 'dependent' variables and 1 constant).

Figure 1

Path analytic representation of one-way MANOVA with 2 groups and 3 dependent variables



Properties of the LR χ^2. It is well known that large samples are needed in order to provide for the valid use of the χ^2 test, the derivation of efficient parameter estimates, and the avoidance of improper solutions (Boomsma, 1983). However, in large samples, the LR χ^2 test may lead to a rejection of a theoretically useful model that is closely aligned with the sample covariance structure (Bollen, 1986, 1989). Similarly, a poor fit based on a small sample size may result in a nonsignificant χ^2. Nevertheless, the exact distribution of estimators and test statistics used in structural modelling is unknown. Monte Carlo studies have investigated the effects of violations of assumptions and the effects of sample size in various models under ML estimation, but this process of simulating artificial data has not covered a wide range of models and conditions. The following 3 sections contain a review of the literature relating to the use of the LR χ^2 statistic under conditions in which traditional MANOVA Type I error rates have been shown to be affected (e.g., inequality of sample n's, data non-normality, inequality of variance-covariance matrices).

Small Sample Behaviour of the LR χ^2 Test. Geweke and Singleton (1980) investigated the small sample behaviour of the LR χ^2 in exploratory factor analysis (FA) models using ML estimation. They simulated a number of FA models with sample sizes of 10, 30, 150, and 300 and found that with small samples (i.e., n $\leq$30) the behaviour of the LR χ^2 test statistic does not follow a χ^2 distribution.

However, the asymptotic distribution theory of the test became valid as sample size increased and the number of factors being fit decreased. They concluded that the LR χ^2 test had considerable power, even in small samples. Unfortunately, the authors only used simple models with a small ratio of number of parameters estimated to number of subjects. Bearden et al. (1982) explored the behaviour of the LR χ^2 test statistic in causal models with sample sizes ranging from 25 to 10,000. The authors simulated and analyzed a number of two construct and four construct SEM's, incorporating ML estimation and multivariate normal data. They found that the overall fit statistic provided by LISREL appeared to be χ^2 distributed for simple models over a wide range of sample sizes. The statistic was not, however, χ^2 distributed for more complex models when the sample size was small.

Boomsma (1982) also studied the robustness of LISREL against small to moderate sample sizes in FA models. Maximum Likelihood estimation was used to test 6 or 8 variable 2 factor models in samples ranging in size from 25 to 400. The results of the analysis indicated that the LR χ^2 test statistic deviated considerably from the theoretical χ^2 distribution in small samples. Furthermore, the robustness of the statistic was related to model complexity. That is, models with larger number of variables and factors may require larger n's. The author recommended that a minimum sample size of 200 be used, and that it may be

dangerous to use sample sizes of less than 100 for the FA models that were studied.

Gallini and Mandeville (1984) examined the validity of the LR χ^2 test in different instances of model misspecification and sample size. They used Monte Carlo methods to study SEM's with 4 latent variables and 7 measured variables in sample sizes ranging from 50 to 500. The simulated data were drawn from a multivariate normal population and analyzed via ML in the LISREL computer program. The results of the study demonstrated that the sensitivity of the LR χ^2 statistic for causal models is linked to the size of the sample, the complexity of the model, and the size of the factor loadings. The authors recommended that other goodness-of-fit (GOF) indices could be used to assess model fit (e.g., the absolute value of the residuals, modification indices).

Gerbing and Anderson (1985) extended Boomsma's (1982) work by studying the effects of sampling error and model characteristics upon parameter estimates and their associated standard errors (SE). They used sample sizes ranging from 50 to 300, drawn from multivariate normal populations, to simulate a number of confirmatory FA models. The number of indicators per factor, number of factors, correlation between factors and indicator reliabilities were all systematically varied and analyzed via ML in the LISREL computer package.

The results indicated that the size of the SE's of the parameter estimates decreased as the sample size increased. Nevertheless, very large samples were not needed in order to achieve robust parameter estimates within the LISREL program. Robust estimates could be obtained in samples with less than 200 cases.

Tanaka (1987) provided a comprehensive review of the Monte Carlo studies that explored the robustness of the LR χ^2 test in small samples. In general, the LR χ^2 test is χ^2 distributed for simple models over a wide range of sample sizes. However, this result may not hold for small samples and complex models. The author concluded that the appropriateness of sample size is linked to the number of parameters estimated in the model and the complexity of the estimation method.

Robustness of the LR χ^2 in Moment Structure Models. The robustness of more traditional statistical procedures against violation of assumptions has often been tested but methods applied to SEM's involving moment structures have only recently been explored. More specifically, the conditions under which the LR χ^2 is valid for testing the equality of means in SEM multiple-group designs has not been sufficiently investigated.

A recent study by Kühnel (1988) showed how programs such as LISREL

can be used to test one-way MANOVA designs. Monte Carlo methods were used to compare the traditional one-way MANOVA analysis (via SPSS-X) and the multiple-group SEM approach in assessing the equality of means between independent groups. Fifty random samples of 500 cases were drawn from multivariate normal populations with 4 variables in each of 4 groups. The equality of means and equality of variance-covariance's were then varied in a 2x2 experimental design. The results indicated that when variance-covariance's were equal among groups both specifications yielded acceptable Type I error rates. When the variance and covariances were not equal the test based on least-squares regression performed poorly, having a Type I error rate (i.e., rejection of a correct null hypothesis) of 12% and a Type II error rate (i.e., not rejecting a false null hypothesis) of 74%. However, the LR test for the model based on group comparisons produced the same Type I error rate, regardless of variance-covariance heterogeneity.

Although the study by Kühnel (1988) was limited in terms of the conditions that were explored, it did show that, in certain circumstances, the LISREL group comparison approach was preferable to the MANOVA test criterion based on least-squares dummy variable regression. The author did note, however, that in SEM all statistical inferences are asymptotic and the results are valid only in moderately large samples. Furthermore, the LR χ^2 test in LISREL is particularly

sensitive to decreasing sample size and model complexity (i.e., ratio of average sample size to the number of dependent variables). The author concluded that testing means with LISREL group comparison should be limited to "one-way MANOVA designs with few dependent variables and few cells where the sample size should not be too small" (Kühnel, 1988, p. 513).

Unfortunately, Kühnel (1988) did not empirically address the issues related to sample size, number of dependent variables, ratio of number of subjects to dependent variables, and the extent of inequality of sample sizes between groups. Although the valid use of the LR χ^2 may be related to the number of dependent variables, groups and cases, the determination of the adequacy of the multiple-group comparisons under suboptimal conditions is essential in ascertaining its utility. Therefore, there would appear to be a need for research aimed at determining the robustness of the LR χ^2 test in multiple-group designs subject to conditions that may yield misleading statistical results (i.e., variance-covariance heterogeneity, inequality of sample sizes, and complex models).

Bagozzi and Yi (1989) expanded on the SEM model proposed by Kühnel (1988) to accommodate other useful designs, including one-way and two-way MANOVA, one-way MANCOVA, stepdown analysis, and models which incorporate multiple measures of one or more criteria. They discussed the

application of SEMs utilizing both the least-squares dummy regression approach as well as the multiple-group approach. The authors noted that, in general, the structural equation approach has a number advantages over traditional multivariate analyses. First, the LR χ^2 in multiple-group models can be used to test experimental effects even if the homogeneity assumption is violated (Kühnel, 1988). The applicability of the test also extends to situations in which the sample sizes are not equal across groups. Second, the procedure provides a way to correct for measurement error in the measures of the variables, thus reducing the chances of making Type II errors. Bagozzi and Yi (1989) concurred with Kühnel (1988) in their assessment of the limitations of this new methodology. That is, as the number of factors and number of measures increase, the number of free parameters to estimate increases, thereby increasing the chances of nonconvergence and improper solutions in estimation.

Effect of Non-normality on the LR χ^2 statistic. Normal distribution theory estimation (i.e., ML) methods become limited information techniques in the presence of non-normal data and can produce bias in GOF indices, including the LR χ^2 statistic (Sharma, Durvasula, & Dillon, 1989). Asymptotically Distribution Free (ADF) estimation methods (e.g., Bentler, 1985; Browne, 1984) can be used in the presence of non-normal data, but these techniques may need considerably larger sample sizes in order to properly estimate 4th order moments.

Browne (1984) found in a simulation study that the χ^2 distribution may be a poor approximation for the distribution of the likelihood ratio χ^2 statistic if the observed variable(s) has (have) substantial kurtosis. Sharma et al. (1989) studied the performance of non-normal estimation methods in the analysis of confirmatory FA models. They conducted a Monte Carlo simulation experiment with a factorial design involving three levels of skewness, three levels of kurtosis, and three different sample sizes. They found that for non-normal conditions the LR χ^2 test statistic for ML and GLS does not follow a χ^2 distribution. In addition, the performance of normal theory methods in non-normal samples was poor and worsened with an increase in kurtosis. The authors recommended that ML not be used when the data come from a non-normal distribution.

Alternatives to the LR χ^2 Test. The main feature of the distributional properties of the ML estimates and the LR χ^2 test of goodness of fit is their asymptotic nature under independent sampling and multivariate normality. However, issues related to model complexity and sample size may be quite varied depending on the estimation strategy employed, the choice of statistic used to measure the goodness-of-fit (GOF) of the model, or the distributional properties of the data. Because of the influence of sample size on the LR χ^2 statistic, researchers have also developed a number of goodness-of-fit tests that are purportedly less sensitive to sample size (e.g., Hoetler, 1983; Jöreskog & Sörbom,

1986; Tucker & Lewis, 1973). A comprehensive evaluation of the use of GOF indices in SEM's can be found in Mulaik, James, Van Alstine, Bennett, Lind and Stilwell (1989).

There have been several recent studies that have addressed the effects of model characteristics and sample size on the parameter estimates and various GOF indices other than the LR χ^2 test statistic. Marsh, Balla and McDonald (1988) used Monte Carlo methods to study the effects of sample size on over 30 GOF indices in CFA models. They found that when the variables to be fit and the model to be tested were held constant, values for most GOF indices were affected by sample size. Furthermore, the effect of sample size tended to be largest for the smallest sample sizes. La Du and Tanaka (1989) also studied the effects of sample size, estimation method, and model misspecification on a number of fit indices. The authors were interested in exploring the extent to which the estimation method used in the solution of SEM's affects the value of GOF indices. The investigators performed a sampling study in which the estimation method, sample size, and model misspecification were systematically varied. They found that, for the same model and data, the average GOF index values for solutions using different estimation strategies might lead to different conclusions. They concluded that sample size and estimation method are important factors in assessing model fit.

CHAPTER 3

Statement of the Problem

The scope of this study, including its purpose, assumptions, and limitations is presented in this chapter. A brief discussion of the significance of this research is also provided.

Justification

Hotelling's T^2, as with the more general MANOVA, requires homogeneity of the variance-covariance matrices between groups. The violation of this assumption in designs with small and/or unequal samples can severely affect the validity of the test statistic (Hakstian et al., 1979). Recently, multiple-group structural equation models have been adapted to incorporate means. As a result, it is possible to use SEM to test for differences in population means between groups. In addition, the resulting LR χ^2 test statistic does not appear to require the variance-covariance homogeneity assumption. Therefore, the comparison of Hotelling's T^2 statistic and the LR χ^2 test statistic under a variety of conditions is an essential process in determining the potential benefits of using SEM to test for differences between group means.

Purpose

The purpose of this study was to compare the Type I error rates for Hotelling's T^2 test and the LR χ^2 test in the analysis of typical one-way (2-group) MANOVA designs. This investigation was accomplished through Monte Carlo methods.

Assumptions and Limitations

1. Since Monte Carlo methods were used, the generalizations of the results are limited to the range of values for the number of dependent variables (p), the absolute sample size, the ratio of sample sizes $(n_1:n_2)$, the ratio of the average sample size to the number of dependent variables $((n_1+n_2)/2/p)$, and the relationship between the variance-covariance matrices. The use multivariate normal data also precludes the generalization of the results to situations involving non-normal data distributions.

2. The conclusions drawn from the study are limited to the extent which the Monte Carlo simulations model the theory. Since 2000 repetitions of each condition were performed, the departure of the expected results from the

actual results should be minimal.

Significance of the Study

The properties of the LR χ^2 test statistic under conditions of sample size, model misspecification, and estimation strategy have been addressed by a number of authors. Likewise, there has been some research aimed at ascertaining the robustness of the SEM group comparison methodology in analyzing traditional MANOVA designs. Given that the SEM group comparison strategy may hold some advantage over the traditional MANOVA test procedures (e.g., Hotelling's T^2, Wilks' Lambda), there remains a need to quantify this benefit. The results of this study will provide evidence of the relative performance of the LR χ^2 test under violation of the homogeneity assumption and relatively small n. Furthermore, by incorporating the SEM test (LR χ^2) for group mean differences under realistic conditions that were previously investigated by Hakstian et al., (1979), information on the behaviour of the LR χ^2 in practical research scenarios will be provided. In addition, the comparative advantages and disadvantages of the LR χ^2 test and Hotelling's two-sample T^2 in these conditions should furnish researchers with some guidelines regarding the choice of a statistic for significance testing.

CHAPTER 4

Methodology

A discussion of the methodology, focusing on the Monte Carlo design and data generation procedures, is presented in the following chapter. The specific conditions under which Hotelling's two-sample T^2 statistic and the LR χ^2 statistic were tested are outlined and explained.

Hakstian et al. (1979) performed a comprehensive investigation of the robustness of Hotelling's two-sample T^2 test with respect to the violation of the homogeneity assumption. The conditions that were modelled represented realistic departures from optimal conditions that would conceivably be found in real world data. Therefore, the present study utilized a subset of identical simulations. This provided a basis for comparison with previous findings regarding Hotelling's T^2 as well as some indication that the simulated data were being generated correctly.

The present study examined the influence of average sample size (n), extent of inequality of sample sizes (n_1:n_2), number of variables (p), and extent of inequality of covariance matrices (Σ's) on the Type I error rates produced by Hotelling's T^2 test and the LR χ^2 test for the two-group p-variate simultaneous

comparison of means. Hotelling's two-sample T^2 was calculated through a FORTRAN subroutine. The LR χ^2 test was based on output from the LISREL VI computer program. The simulated data were derived from multivariate normal populations. The general framework was similar to that incorporated by Hakstian et al. (1979) with only minor modifications. The study utilized the following manipulations:

1. The ratio of average sample size $(n_1+n_2)/2$ to the number of variables. Two ratios were utilized: 10 subjects per variable and 30 subjects per variable. Hakstian et al. (1979) employed a ratio of 3 subjects per variable as one of their qualifications. In the present study this condition was not employed. In general, small samples are not compatible with maximum likelihood estimation, depending on the complexity of the model and the population moment matrices to be analyzed. Therefore, the utilization of an extremely small number of subjects per variable would likely result in indeterminant solutions in the structural equation model.

2. The inequality of sample sizes. The following ratios of the sizes of the two samples were implemented: 1:1, 2:1, and 5:1. It should be noted that when $n_1 \neq n_2$ and $\Sigma_1 \neq \Sigma_2$ it would make a difference whether the larger sample was drawn from the population with larger dispersions or smaller dispersions. When $n_1 = n_2$ this

does not apply. Therefore, for each condition of $\Sigma_1 \neq \Sigma_2$ (to be discussed later) there were 5 possible conditions relating to the inequality of sample sizes.

3. The number of variables. Two levels were employed: 2 and 6. The utilization of larger numbers of variables was not suggested in that, within the SEM framework, it would require the estimation of more free parameters, resulting in the consumption of a large number of degrees of freedom. Therefore, the derivation of ML estimates may be problematic. It should also be noted that the combination of $(n_1+n_2)/2:p=10:1$ and $p=2$ would result in a total sample size of 40. Although a total sample size of this magnitude may be considered *a priori* "too small" for latent variable modelling (e.g., Boomsma, 1982), an examination of the utility of the LR χ^2 in moment structure models of this size is warranted in practice.

4. The heterogeneity of covariance matrices. Four conditions were implemented: (1) The variables in population 1 were set with σ's = 1 and the variables in population 2 were rescaled by 1.2. As a result, the variances in population 2 were 1.44x those in population 1. This was viewed as a mild departure from homogeneity and was referred to as heterogeneity condition 1; (2) The variables in population 1 were set with σ's = 1 and the variables in population 2 rescaled by 1.5. This resulted in the variables in population 2 having variances that were

2.25x those in population 1. This was viewed as a moderate departure from homogeneity and was referred to as heterogeneity condition 2; (3) The first p/2 variables in population 2 were rescaled by 1.5 and the remaining p/2 variables were left with $\sigma=1$. This resulted in a condition where heterogeneity was present on some but not all variables. This was referred to as heterogeneity condition 3; (4) The variables were left on the same scale ($\Sigma_1=\Sigma_2$). This was referred to as the homogeneity condition.

Similar to Hakstian et al. (1979), an attempt was made to generate data that would possess the same heterogeneity often found in real life problems. The population 1 covariance matrices (Σ_1) for the p=2 and p=6 levels were identical to those used by Hakstian et al. (1979). These matrices are presented in Table 1.

The population 2 covariance matrices (Σ_2) for each value of p were derived from $\Sigma_2=D_i\Sigma_1D_i$, where i=1,2,3,4. The diagonal D_i matrices contained the factors mentioned above. For heterogeneity conditions 1 and 2 the diagonal matrices D_1 and D_2 contained 1.2 and 1.5 in the diagonal positions, respectively. For heterogeneity condition 3, D_3 contained 1.5 in the first p/2 diagonal elements and 1 in the final p/2 positions. For the homogeneity condition, D_4 was simply the identity matrix.

Table 1

Population Covariance Matrices Used as Σ_1

2 Variables			6 Variables						
Variable	1	2	Variable	1	2	3	4	5	6
1	1.00		1	1.00					
2	.50	1.00	2	.40	1.00				
			3	.30	.35	1.00			
			4	.40	.25	.10	1.00		
			5	.30	.40	.50	.30	1.00	
			6	.20	.30	.00	.40	.15	1.00

Note

The Σ_2 matrices used in this study were derived from transformations of the matrices in Table 1. The process of rescaling is described in the text.

These manipulations resulted in the examination of 72 unique conditions. There are 60 (2x5x2x3) conditions where $\Sigma_1 \neq \Sigma_2$ and 12 (2x3x2) conditions where $\Sigma_1 = \Sigma_2$. The GGNSM subroutine from the International Mathematical and Statistical Library (IMSL, 1982) was used to generate multivariate random normal deviates with a given covariance structure dictated by Σ_1 and Σ_2. For each of the 72 conditions, 2000 pairs of samples from populations in which the null hypothesis was true were generated, resulting in the simulation of a total of 144,000 experiments. For each pair of samples a T^2 statistic was calculated and transformed to an F value. The F value was then compared to the 90th, 95th and 99th percentile points of the appropriate F distribution. Tabulations were then made as to the number of instances that the actual Type I error rate exceeded the nominal value for $\alpha = .10$, .05, and .01, over the 2000 repetitions. The generated data were then input to a LISREL VI program which tested for the equivalence of mean structures between the independent samples. The LISREL routine produced a LR χ^2 statistic and accompanying p value which was used to ascertain the actual Type I error rates across conditions. This p value indicates the likelihood of observing a χ^2 value of this magnitude or larger, assuming that the specified model was correct. A contrast of the actual Type I error rates for the two tests under the aforementioned conditions was then undertaken. The percentage of statistically significant results (those that exceed the nominal α levels) for each condition was compared for the two test statistics (Hotelling's T^2 and the LR χ^2).

CHAPTER 5

Results

The summary of the results is presented in the following chapter. Specific comparisons between Hotelling's two-sample T^2 statistic and the LR χ^2 statistic are outlined and interpreted. In addition, a contrast between the likelihood ratio χ^2 statistic and some of the alternatives to Hotelling's two-sample T^2 (i.e., Yao's test) is provided.

The actual Type I error rate (τ) was calculated for each of the two test statistics in each of the 72 conditions. The proportion of results, based on 2000 samples, that exceeded the nominal α levels ,.01, .05 and .10, are reported in Tables 2, 3, 4, and 5. Because of fluctuations due to sampling, it is necessary to set limits on the expected Type I error rates. Therefore, approximate 95% confidence intervals were formed using the standard error of a proportion and normal curve probabilities. If $|\tau - \alpha| < 2[\alpha(1 - \alpha)/2000]^{1/2}$ then that value of τ could be considered to be within sampling error of its respective nominal value.

Table 2

Actual Type I Error Rates for the Two-Sample T^2 Statistic and the Likelihood Ratio χ^2 for p=2 and $(n_1 + n_2)/2/p = 10$.

				Heterogeneity 1		Heterogeneity 2		Heterogeneity 3	
n1:n2	Test	α	Homo-geneity	Pos	Neg	Pos	Neg	Pos	Neg
20:20	T^2	.01	.009	.005		.010		.005	
	LR χ^2		.009	.006		.010		.006	
	T^2	.05	.051	.039		.049		.048	
	LR χ^2		.051	.042		.052		.051	
	T^2	.10	.096	.088		.102		.092	
	LR χ^2		.100	.090		.105		.094	
27:13	T^2	.01	.007	.004	.015	.002	.029	.007	.014
	LR χ^2		.011	.009	.011	.008	.011	.009	.008
	T^2	.05	.048	.029	.064	.021	.086	.025	.063
	LR χ^2		.046	.040	.048	.039	.047	.039	.038
	T^2	.10	.085	.060	.117	.041	.157	.052	.117
	LR χ^2		.089	.084	.098	.075	.095	.087	.086
33:7	T^2	.01	.009	.005	.025	.000	.067	.004	.032
	LR χ^2		.016	.010	.016	.007	.018	.017	.011
	T^2	.05	.058	.027	.093	.005	.168	.037	.097
	LR χ^2		.066	.056	.077	.056	.073	.060	.056
	T^2	.10	.112	.059	.173	.019	.262	.077	.164
	LR χ^2		.126	.115	.129	.104	.129	.124	.114

Note

The degree of heterogeneity is explained in the text
Positive (POS) Condition: The population whose covariance matrix contains the larger entries is that from which the larger n was drawn.
Negative (NEG) Condition: The population whose covariance matrix contains the smaller entries is that from which the larger n was drawn.

Table 3

Actual Type I Error Rates for the Two-Sample T^2 Statistic and the Likelihood Ratio χ^2 Statistic for p=2 and $(n_1 + n_2)/2/p = 30$.

n1:n2	Test	α	Homo-geneity	Heterogeneity 1 Pos	Heterogeneity 1 Neg	Heterogeneity 2 Pos	Heterogeneity 2 Neg	Heterogeneity 3 Pos	Heterogeneity 3 Neg
	T^2	.01	.005	.004		.009		.005	
	$LR\chi^2$		.005	.004		.008		.005	
60:60	T^2	.05	.043	.045		.046		.043	
	$LR\chi^2$		.044	.045		.046		.044	
	T^2	.10	.090	.086		.090		.098	
	$LR\chi^2$		.091	.086		.091		.099	
	T^2	.01	.008	.005	.016	.025	.004	.012	
	$LR\chi^2$		.008	.010	.007	.007	.007	.007	.006
80:40	T^2	.05	.041	.040	.057	.015	.094	.029	.069
	$LR\chi^2$		.043	.052	.042	.036	.050	.041	.045
	T^2	.10	.086	.081	.111	.037	.165	.057	.126
	$LR\chi^2$		.088	.107	.087	.088	.099	.084	.088
	T^2	.01	.015	.004	.028	.001	.056	.005	.032
	$LR\chi^2$		.013	.009	.009	.009	.013	.007	.010
100:20	T^2	.05	.053	.024	.100	.007	.159	.025	.108
	$LR\chi^2$		.052	.058	.056	.047	.047	.042	.049
	T^2	.10	.100	.051	.171	.021	.236	.052	.179
	$LR\chi^2$		.103	.105	.114	.092	.097	.092	.098

Note

The degree of heterogeneity is explained in the text
Positive (POS) Condition: The population whose covariance matrix contains the larger entries is that from which the larger n was drawn.
Negative (NEG) Condition: The population whose covariance matrix contains the smaller entries is that from which the larger n was drawn.

Table 4

Actual Type I Error Rates for the Two-Sample T^2 Statistic and the Likelihood Ratio χ^2 Statistic for p=6 and $(n_1 + n_2)/2/p = 10$.

n1:n2	Test	α	Homogeneity	Heterogeneity 1 Pos	Heterogeneity 1 Neg	Heterogeneity 2 Pos	Heterogeneity 2 Neg	Heterogeneity 3 Pos	Heterogeneity 3 Neg
60:60	T^2	.01	.011	.012		.011		.008	
	$LR\chi^2$		.013	.013		.011		.009	
	T^2	.05	.052	.052		.047		.043	
	$LR\chi^2$		.056	.057		.052		.052	
	T^2	.10	.096	.100		.097		.094	
	$LR\chi^2$		.108	.109		.108		.103	
80:40	T^2	.01	.009	.001	.014	.003	.033	.006	.023
	$LR\chi^2$		.011	.006	.008	.014	.011	.013	.011
	T^2	.05	.044	.032	.068	.014	.122	.031	.082
	$LR\chi^2$		.046	.057	.044	.054	.056	.061	.061
	T^2	.10	.097	.068	.136	.033	.204	.069	.155
	$LR\chi^2$		.112	.115	.098	.113	.109	.119	.108
100:20	T^2	.01	.009	.002	.043	.000	.122	.002	.052
	$LR\chi^2$		.019	.017	.028	.015	.028	.015	.020
	T^2	.05	.044	.015	.118	.002	.256	.012	.158
	$LR\chi^2$		.076	.069	.089	.061	.091	.063	.085
	T^2	.10	.093	.034	.204	.006	.363	.037	.237
	$LR\chi^2$		.133	.131	.155	.112	.171	.118	.156

Note

The degree of heterogeneity is explained in the text
Positive (POS) Condition: The population whose covariance matrix contains the larger entries is that from which the larger n was drawn.
Negative (NEG) Condition: The population whose covariance matrix contains the smaller entries is that from which the larger n was drawn.

Table 5

Actual Type I Error Rates for the Two-Sample T^2 Statistic and the Likelihood Ratio χ^2 Statistic for p=6 and $(n_1 + n_2)/2/p = 30$.

				Heterogeneity 1		Heterogeneity 2		Heterogeneity 3	
n1:n2	Test	α	Homo-geneity	Pos	Neg	Pos	Neg	Pos	Neg
180:180	T^2	.01	.007	.011		.010		.005	
	$LR\chi^2$		.007	.011		.010		.005	
	T^2	.05	.044	.047		.045		.049	
	$LR\chi^2$		.044	.049		.046		.050	
	T^2	.10	.087	.099		.090		.098	
	$LR\chi^2$		.090	.103		.093		.103	
240:120	T^2	.01	.009	.003	.019	.002	.039	.002	.024
	$LR\chi^2$		.008	.009	.007	.010	.013	.008	.008
	T^2	.05	.055	.023	.071	.013	.136	.026	.096
	$LR\chi^2$		.055	.049	.043	.051	.048	.046	.048
	T^2	.10	.105	.058	.146	.030	.227	.056	.154
	$LR\chi^2$		.112	.088	.099	.105	.100	.101	.109
300:60	T^2	.01	.011	.002	.033	.000	.134	.004	.056
	$LR\chi^2$		.012	.012	.011	.010	.014	.013	.009
	T^2	.05	.051	.015	.125	.001	.278	.018	.163
	$LR\chi^2$		.055	.057	.054	.055	.063	.058	.053
	T^2	.10	.096	.031	.210	.006	.390	.049	.252
	$LR\chi^2$		.103	.117	.117	.116	.113	.113	.108

Note

The degree of heterogeneity is explained in the text
Positive (POS) Condition: The population whose covariance matrix contains the larger entries is that from which the larger n was drawn.
Negative (NEG) Condition: The population whose covariance matrix contains the smaller entries is that from which the larger n was drawn.

There is no generally accepted, standard, quantitative definition of what constitutes robustness. Bradley (1978) suggests that a definition of robustness should make the robustness criterion proportional to α. He defines a liberal criterion of robustness as one in which $|\tau-\alpha| \leq \alpha/2$. Algina and Tang (1988) proposed that conditions in which τ was outside of the 95% confidence interval for the expected Type I error rate for at least 2 of α=.01, .05, and .10 could be classified as being non-robust. The more conservative criterion of Algina and Tang is employed in the interpretation of the present results.

Hotelling's T^2

As expected, the findings regarding the Type I error rates based on Hotelling's T^2 were similar to those of Hakstian, et al.. Using the Algina and Tang (1988) criterion, it is clear that under conditions of variance-covariance homogeneity and/or equal sample sizes, the Type I error rates fall within the nominal bounds (see Tables 2-5). However, the combination of unequal sample sizes and between-group variance-covariance heterogeneity almost always yielded Type I error rates that were atypical of the nominal levels, regardless of the degree of heterogeneity and/or inequality of the sample sizes. The severity of the departures of τ from α were influenced by several factors.

First, the degree of departure of τ from α escalated as the ratio n_1:n_2 departed from 1. That is, if the degree of heterogeneity and the $(n_1+n_2)/2$:p ratio were held constant, the severity of the departure between the actual and nominal error rates escalated as the inequality of the sample sizes increased. Furthermore, when the larger sample was drawn from the population whose variance-covariance matrix contained the larger values, the actual Type I error rate tended to be conservative. In contrast, when the larger sample was drawn from the population whose variance-covariance matrix contained the smaller values, the actual Type I error rate tended to be liberal. Second, in general, the disparity between τ and α noted above was amplified by an increase in the overall sample size. That is, for a fixed n_1:n_2 ratio, degree of heterogeneity and p, the discrepancy between τ and α tended to be more pronounced for the larger average sample size. This is evident when one compares the results of Table 4 with Table 5 but is less apparent when comparing Table 2 and Table 3. Finally, when the degree of heterogeneity, the average sample size and the degree of inequality of sample sizes were all held constant, the more sizable differences between τ and α were generally associated with the smaller average sample size to dependent variable ratio ($(n_1+n_2)/2$:p=10:1). This is evident when one compares the results in Tables 3 and 4.

Likelihood Ratio χ^2

The results pertaining to the LR χ^2 were quite different from those for Hotelling's T^2. The LR χ^2 performed favourably when the sample sizes were equal, regardless of the ratio of the average sample size to the number of dependent variables or the degree of variance-covariance heterogeneity. The Type I error rates in 11 of the 12 possible conditions were within the bounds of the criterion set by Algina and Tang.

In the cases where $\Sigma_1=\Sigma_2$, the LR χ^2 statistic yielded Type I error rates that were within nominal levels in all situations except when $n_1:n_2=5:1$ or 1:5 and the ratio of average sample-size to number of dependent variables, $(n_1+n_2)/2:p$, was 10:1 (Tables 2 and 4). There were no meaningful departures of τ from α when $(n_1+n_2)/2:p=30:1$ (Tables 3 and 5).

When $\Sigma_1\neq\Sigma_2$, several notable trends were apparent. From Tables 3 and 5, it is evident that the LR χ^2 produced Type I error rates that fell within the nominal bounds (as set out by Algina and Tang, 1988) in 29 of the 30 conditions where $(n_1+n_2)/2:p=30:1$. The only exception was found in Table 3 where, with equal sample sizes and mild heterogeneity (Heterogeneity 1), the Type I error rate was conservative. The results were quite different, however, when $(n_1+n_2)/2:p=10:1$. In these cases (Tables 2 and 4), the Type I error rates were within sampling error

of their respective nominal levels in all 6 cases where the sample sizes were equal between groups. However, when $n_1:n_2=2:1$ or 1:2, the number of Type I errors differed from the expected nominal levels in 3 of the 12 cases. When $n_1:n_2=5:1$ or 1:5, there were 9 out of 12 conditions in which the actual Type I error rates deviated from their nominal levels. These results indicate that the degree of inequality in sample sizes affects the Type I error rates at the $(n_1+n_2)/2:p=10:1$ ratio.

The overall utility of the LR χ^2 test would appear to be dependent on the ratio of the average sample size to the number of dependent variables. An examination of Tables 3 and 4, in which the sample sizes are constant (average $n=60$) but the number of dependent variables differ ($p=2$ and $p=6$), clearly indicates that as the ratio $(n_1+n_2)/2:p$ increases the departure of τ from α decreases.

Likelihood Ratio χ^2 vs Hotelling's T^2

A number of noteworthy trends emerge from the direct comparison of the results derived from the two test statistics. First, both test statistics produce appropriate Type I error rates when $n_1:n_2=1:1$ regardless of the sample size, the number of dependent variables or the degree of variance-covariance heterogeneity.

Second, important differences between the two techniques are apparent when $\Sigma_2=\Sigma_1$. Hotelling's T^2, as expected, yields nominal Type I error rates in all conditions. The LR χ^2 did, however, produce inflated Type I error rates in the condition where the ratio of average sample size to the number of dependent variables was 10:1 and the ratio of n_1:n_2 was 5:1. This suggests that with very unequal sample sizes, and a small average sample size to dependent variable ratio, Hotelling's T^2 statistic would be preferred for the test of group mean differences. However, due to the limitations of the ratios selected in this study, it is difficult to provide more specific guidelines as to the exact conditions under which one statistic would be preferable to the other.

Third, when sample sizes were large and unequal, combined with any degree of between-group variance-covariance heterogeneity, the results produced by the LR χ^2 statistic and Hotelling's T^2 statistic were remarkably dissimilar. In all of these cases, Hotelling's T^2 produced unacceptable Type I error rates. In contrast, when $(n_1+n_2)/2$:p was 30:1 the number of rejections of the null hypothesis using the LR χ^2 was almost always within sampling error of its' nominal value. Although this result did not always hold under the lower average sample size to dependent variable ratio ($(n_1+n_2)/2$:p=10:1), the degree of departure from the nominal level was much less for the LR χ^2 test statistic than for Hotelling's T^2 test in all conditions.

Fourth, when $n_1 \neq n_2$ and the homogeneity assumption was violated, the validity of Hotelling's T^2 was suspect. In general, there was an inflation of the Type I error rate when the group with the larger sample size had the larger variances. In contrast, a conservative test statistic was produced when the smaller group had the larger variances. This trend was not evident for the LR χ^2 statistic. More specifically, the Type I error rates for the LR χ^2 statistic were not systematically affected by the relative size of the two samples in relation to variance-covariance heterogeneity.

Finally, for a fixed $n_1:n_2$ ratio and degree of variance-covariance heterogeneity, the departure of τ from α tended to be more pronounced for larger average sample sizes when the T^2 statistic was used. In contrast, if the $n_1:n_2$ ratio and the degree of heterogeneity were held constant, the LR χ^2 statistic became more robust with an increase in average sample size.

Likelihood Ratio χ^2 vs Yao's and James' tests

Yao's approximate degrees of freedom test and James' first order asymptotic series test have both been suggested for use when the homogeneity assumption is violated. Algina and Tang (1988) report that both of these tests are less sensitive to covariance heterogeneity than is Hotelling's T^2. The authors also report that Yao's test was superior to James' test in terms of its ability to provide

actual Type I error rates that were close to the nominal level. Therefore, the following comparisons will involve only the LR χ^2 test and Yao's test.

Algina and Tang (1988) found that Yao's test had appropriate Type I error rates when $p \leq 10$, $(n_1+n_2)/p \geq 10$, $1:2 \leq n_1:n_2 \leq 2:1$ and $\Sigma_2 \leq 9\Sigma_1$. The simulated conditions for the LR χ^2 that were within these limits generally yielded robust results, except when $p=2$, $(n_1+n_2)/p=20$, and $n_1:n_2=2:1$ or 1:2. An inspection of the results for this set of outcomes reveals that, although 2 of the conditions resulted in $|\tau - \alpha| \geq 2[\alpha(1 - \alpha)/2000]^{1/2}$ for at least 2 of $\alpha=.01, .05, .10$, the actual deviations of τ from α were not extreme (see Table 2). In fact, the mean Type I error rates for these runs were .010, .042, and .088 for $\alpha=.01, .05, .10$, respectively.

Algina and Tang (1988) concluded that if $(n_1+n_2)/p \geq 20$, Yao's test is robust when a) $n_1:n_2$ is 5:1 or smaller and $p=2$ or b) $n_1:n_2$ is 3:1 or smaller and $p=6$. The results of the present study indicate that when $(n_1+n_2)/p=20$, $p=2$, and $n_1:n_2=5:1$ or 1:5 the use of the LR χ^2 results in a mildly inflated Type I error rate. Nevertheless, an inspection of Table 2 reveals that the discrepancies between α and τ are not severe. For $\alpha=.01, .05, .10$, the mean Type I error rates were .014, .064, and .120, respectively. Unfortunately, the present study did not consider conditions in which $n_1:n_2=3:1$. Therefore, it was impossible to directly assess the appropriateness of the LR χ^2 test under the previously defined limits for which Yao's test has been found to be robust. However, when $(n_1+n_2)/p=20$,

$p=6$, and $n_1:n_2=2:1$ or 1:2, the LR χ^2 test produced generally acceptable Type I error rates (see Table 3). Although one of the runs produced results which were outside of sampling error for 2 of $\alpha=.01, .15, .10$, the robustness of the LR χ^2 statistic under these conditions is clearly evident.

Algina and Tang (1988) also reported Type I error rates for $\alpha=.05$, $\Sigma_2=D_i\Sigma_1D_i$, $n_1\geq n_2$ and $p=2,6$. For $(n_1+n_2)/p=20$, $p=2$, $D_i=1.5$ and $n_1:n_2=20:20$, 27:13 and 33:7 they report actual Type I error rates of .049, .054, and .052. An identical condition in the present study resulted in Type I error rates of .052, .047 and .073 for the LR χ^2 test. For the above conditions, the error rates for Yao's test and the LR χ^2 test are comparable, except when $n_1:n_2=5:1$. For $(n_1+n_2)/p=20$, $p=6$, $D_i=1.5$ and $n_1:n_2=60:60$, 80:40 and 100:20, the estimated Type I error rates for Yao's test were .046, .072 and .064 for $\alpha=.05$. The latter two values were associated with conditions in which $|\tau - \alpha| \geq 2[\alpha(1 - \alpha)/2000]^{1/2}$ for at least 2 of $\alpha=.01, .05, .10$. An identical condition in the present study produced Type I error rates of .052, .056 and .091, respectively. For these results, only the final value was associated with a condition in which $|\tau - \alpha| \geq 2[\alpha(1 - \alpha)/2000]^{1/2}$ for at least 2 of $\alpha=.01, .05, .10$.

CHAPTER 6

Discussion

The utility of the SEM methodology for testing differences in group means is discussed in following section. In addition, the results of the Monte Carlo simulations are related to earlier findings regarding the robustness of Hotelling's T^2 test and the LR χ^2 test. Finally, possible future applications of multiple-group SEM are examined.

Multivariate Analysis of Variance (MANOVA) is a generalization of analysis of variance which allows the researcher to analyze more than one dependent variable at time. This process of comparing group mean differences on several variables simultaneously has been applied to a wide variety of research problems, including numerous studies in education. Recently, there has been a trend towards understanding MANOVA in terms of a general linear model and its relationship to a structural equation model that underlies the technique (Bray & Maxwell, 1982, 1985). This formulation of MANOVA as a special case of SEM has not received widespread attention. Nevertheless, the use of causal modelling in experimental research does offer some distinct statistical and theoretical advantages over more traditional methodologies. Above all it provides a means by which to formulate experiments in terms of unobservable causal processes.

The present study compared the relative effectiveness of the LR χ^2 (SEM test criterion) and Hotelling's two-sample T^2 statistic (MANOVA test criterion) in testing for group mean differences. In conditions where the homogeneity assumption was not violated, the results indicated that the LR χ^2 test may not be appropriate when the disparity in sample sizes between groups is large. However, the LR χ^2 seemed to perform poorly only when the ratio of the average sample size to the number of dependent variables was relatively small (i.e., 10:1). This finding lends support to previous research that addressed the limitations of the SEM approach in testing for group mean differences (Bagozzi & Yi, 1989; Kühnel, 1988). That is, as the model to be tested becomes more complex (i.e., increasing the number of dependent variables, increasing the number of groups) and/or the overall sample size decreases, the greater the chances for improper solutions and a resulting bias in the LR χ^2 statistic. This trend suggests that if a test for the heterogeneity of variance-covariance matrices between groups fails to reject the null hypothesis, Hotelling's T^2 should be used if the ratio of the average sample size to the number of dependent variables is small.

The results also showed that when there are large sample sizes relative to the number of dependent variables, and the homogeneity assumption is not violated, either test is appropriate. This result supports earlier findings regarding the robustness of Hotelling's T^2 statistic under variance-covariance homogeneity (e.g., Hakstian et al., 1979; Ito & Schull, 1964). In addition, the questionable

appropriateness of the LR χ^2 in small samples is supported (e.g., Boomsma, 1982; Geweke & Singleton, 1980). Under violation of the homogeneity assumption, the use of the LR χ^2 test statistic appears to be preferred. In all cases, it controls for the Type I error rate better than Hotelling's T^2 test. This is not surprising in that the multiple-group SEM approach to testing means does not assume homogeneity of variance-covariance matrices. Conversely, the violation of the homogeneity assumption has been found to have a significant effect on the Type I error rates produced by Hotelling's T^2 statistic (e.g., Hakstian et al., 1979). This implies that the SEM approach can provide a more appropriate test of mean differences when the homogeneity assumption is violated. Nevertheless, with an average sample size to dependent variable ratio that is quite small, there is still some inflation in the number of false rejections of the null hypothesis when the LR χ^2 statistic is used. In such cases, other adjustment techniques (e.g., Yao, 1965) might be employed.

There are certain considerations that need to be addressed with respect to the appropriateness of either technique (traditional MANOVA or SEM group means comparison) for a given research problem and data set. These issues will be discussed in the remainder of the section.

In moment structure modelling applied to testing mean differences, the limitations on the sample size in each group and the number of dependent

variables that can be tested is related to the general restrictions inherent in the SEM methodology. First, it is well known that ML estimation and the LR χ^2 statistic apply to relatively large samples. Furthermore, computation problems during optimization are an inverse function of sample size. The required sample size varies, depending on the complexity of the model and the population moment structure in each group. Bagozzi (1981) found that in many instances, ML estimation in covariance structure models is justifiable when the sample size minus the number of parameters to be estimated is greater than 50. In terms of the present analysis, conditions in which the total sample size minus the number of parameters to be estimate was greater than 50 generally yielded acceptable Type I error rates for the LR χ^2 statistic. However, studies with artificial data and models have only recently been applied to utilization of SEMs in these types of experimental designs (e.g., Kühnel, 1988).

The present study investigated the behaviour of the probability of a Type I error under various conditions for Hotelling's T^2 statistic and the LR χ^2 statistic. Therefore the findings are idiosyncratic to the model characteristics that were studied (a one-way, 2-group, comparison of means with 2 and 6 dependent variables). Higher-way designs (e.g., two-way MANOVA) can be formulated with multiple- group SEM approach by expressing one of the independent variables as a dummy variable separately for each group (see Bagozzi & Yi, 1989). However, the comparative advantage of the SEM approach in these more complex designs

has not been studied. It may be difficult to obtain ML solutions in SEMs with numerous dependent variables, groups, and within subject factors. These models would incorporate a large number of free parameters that would need to be estimated and, therefore, would likely require larger n's in order to provide for valid test statistics. Further research in this area is necessary in order to ascertain the relationship between the complexity of the multiple-group model and the minimum sample size needed to provide a valid goodness-of-fit test statistic.

Traditional MANOVA assumes homogeneity of variance-covariance matrices. In the present study Hotelling's T^2 statistic was shown to be relatively robust to the violation of this assumption when $n_1 = n_2$. However, when $n_1 \neq n_2$ there were significant deviations between the expected and actual Type I error rates. In the multiple-group SEM approach to testing group mean differences the homogeneity assumption can be assessed, and proper tests applied even when it is violated. Since the variances and covariances of each group are estimated separately, the homogeneity assumption can be effectively ignored. This will be attractive in many educational research settings where the homogeneity assumption is unrealistic. For example, in studies of attitudes and opinions the homogeneity assumption may not be realistic across subsets of the group being studied (Muthén, 1989). Therefore, the analysis of data as if they were obtained from a single population may not be valid with traditional MANOVA statistics.

Multivariate Analysis of Variance also assumes that the observed variables and errors have multivariate normal distributions. For SEM, the estimation method most often used by empirical researchers is ML. Under conditions of normality (and relatively large samples) ML provides estimates that are consistent and efficient. In addition, the minimization function for ML has an asymptotic χ^2 distribution which allows for the calculation of a LR χ^2 goodness-of-fit statistic. If the assumption that all of the observed variables have multivariate normal distributions is not met, the utility of the LR χ^2 test statistic may be suspect (Sharma et al., 1989). Therefore, in the presence of non-normality, the SEM approach to testing means (using ML) may yield little statistical advantage over traditional MANOVA techniques.

Asymptotically Distribution Free (ADF) estimators are available (Bentler, 1985; Browne, 1984) and could possibly be used in moment structure models. This estimation strategy does, however, require a formidable amount of computation. When data are normally distributed they can be summarized completely in terms of their first, second, and joint moments. By including moments of the 4th order one needs considerably larger samples to provide stable parameter estimates. Nevertheless, the SEM approach to testing mean differences (utilizing ADF estimators) may allow for a relaxation of the normality assumption, depending on the size of the samples in each group.

There are also some practical instances where the choice of a statistical methodology is constrained by the research problem of interest. One strength of SEM is that it allows one to perform tests using latent variables that can be indicated by one or more manifest variables. Many of the variables in behavioral research are unobservable (e.g., anxiety, attitudes, intelligence) but may be theoretically indicated by two or more measures. Therefore, there exists a need for an analytical strategy that takes into account multiple measures of constructs. Unlike traditional MANOVA, SEM provides the methodology to statistically test for the difference in "construct" means between groups.

The SEM approach for testing group mean differences may also be advantageous in terms of post hoc analyses after a significant omnibus test. SEM can be used to test whether the group difference on a certain dependent variable is due to a direct association with the manipulation or to its dependence on other dependent variables. LISREL provides modification indices (MI) that measure the expected decrease in the LR χ^2 statistic if a single constraint is relaxed and the estimated parameters are held fixed at their estimated values. These MIs are approximately distributed as χ^2 with 1 degree of freedom (Jöreskog & Sörbom, 1986) and can be used to determine the dependent variables on which significant mean differences occur. Each time a parameter is freed (e.g., λ_{ik}, which is the mean for the kth dependent variable) based on the MI the LR χ^2 statistic drops in an amount equal to or exceeding the MI. The difference between models in

which this parameter is held constant and one in which it is freed would provide the equivalent of a post-hoc univariate test of the significance of that dependent variable. Because these two models are nested, the difference in their χ^2 statistics, and associated degrees of freedom, can be used to test for the significance of the parameter (mean) that is being freed. Given that general moment structure models can be used to study multiple populations or groups of subjects in experimental settings, it is necessary to explore possible techniques that will allow for appropriate post-hoc analyses. The use LISREL MI may be suitable but additional studies are needed in order to evaluate it's utility.

Finally, one must consider the issue of missing data when using SEMs to test for mean differences. In traditional MANOVA this problem is often countered by deleting the cases with missing data or by calculating a missing value correlation matrix. Within the SEM framework, the calculation of a moment matrix is problematic with missing data. LISREL provides an option for dealing with missing data but this strategy may bias the final χ^2 goodness-of-fit statistic. Until the effect of missing data on the LR χ^2 can be more adequately explored, it is suggested that multiple-group SEMs only be applied in experimental settings where complete data sets are available.

CHAPTER 7

Conclusion

For both the LR χ^2 (SEM test criterion) and Hotelling's two-sample T^2 statistic (MANOVA test criterion), the behaviour of the probability of a Type I error (α) under non-optimal conditions is quite complex and has been shown to be dependent on a number of interacting factors. These factors, which may apply to varying degrees for both test statistics, include the size of α, the absolute sample size, the relative sample size in each group, the absolute shape and variance of the population from which the sample was drawn, and the relative shape and variance of the population from which the sample was drawn.

The present study compared the Type I error rates for Hotelling's T^2 statistic (MANOVA test criterion) and the LR χ^2 statistic (SEM test criterion) across a number of the factors mentioned above. Although several conclusions can be drawn, it should be remembered that the results are idiosyncratic to the specific simulations that were performed. Therefore, no attempt is made to generalize beyond these specific conditions.

First, in terms of maintaining Type I error rates that are close to the nominal levels, the LR χ^2 statistic and Hotelling's T^2 statistic are comparable

when the homogeneity assumption is not violated and the ratio of the average sample size to the number of dependent variables is large (i.e., 30:1). However, when this ratio is small (i.e., 10:1) Hotelling's T^2 statistic is preferred. Second, when the homogeneity assumption is violated, the LR χ^2 statistic provides more appropriate Type I error rates than does Hotelling's T^2 statistic. Although the LR χ^2 statistic does produce an inflated Type I error rate when the ratio of average sample size to the number of dependent variables is small, the number of false rejections of the null hypothesis remains less severe than for Hotelling's T^2 statistic.

The results of this study lend some optimism for the use of the LR χ^2 statistic in testing for group mean differences on a number of dependent variables. The SEM model, and accompanying GOF test (LR χ^2), appear especially promising for cases where the homogeneity assumption is violated. There remains a need, however, for more extensive research aimed at exploring the exact conditions under which this new technique is or is not valid.

REFERENCES

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.

Alwin, D. F. & Tessler, R. C. (1974). Causal models, unobserved variables, and experimental data. American Journal of Sociology, 80, 58-86

Alwin, D. F. & Tessler, R. C. (1985). Causal models, unobserved variables, and experimental data. In H.M. Blalock, jr. (Ed.) Causal models in the social sciences. New York: Aldine.

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

Bagozzi, R. P. (1977). Structural equation models in experimental research. Journal of Marketing Research, 14, 209-226.

Bagozzi, R. P., Fornell, C. & Larcker, D. F. (1981). Canonical correlation analysis as a special case of a structural relations model. Multivariate Behavioral Research, 16, 437-454.

Bagozzi, R. P. & Yi, Y. (1989). On the use of structural equation models in experimental designs. Journal of Marketing Research, XXVI(August 1989), 271-284.

Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. Journal of Marketing Research, 19, 425-430.

Bentler, P. M. (1980). Multivariate analysis with latent variables: Causal modelling. Annual Review of Psychology, 31, 419-56.

Bentler, P. M. (1982). Confirmatory factor analysis via noniterative estimation: a fast, inexpensive method. Journal of Marketing Research, November, 417-424.

Bentler, P. M. (1983a). Some contributions to efficient statistics in structural models: specification and estimation of moment structures. Psychometrika, 48, 493-517.

Bentler, P. M. (1983b). Simultaneous equation systems as moment structure models: with an introduction to latent variable models. Journal of Econometrics, 22, 13-42.

Bentler, P. M. (1985). Theory and implementation of EQS: A structural equation program. Los Angeles: BMDP Statistical Software.

Bentler, P. M. & Weeks, D. G. (1979). Interrelations among models for the analysis of moment structures. Multivariate Behavioral Research, 14, 169-186.

Bentler, P. M. & Weeks, D. G. (1980). Linear structural equations with latent variables. Psychometrika, 45, 289-308.

Bollen, K. A. (1986). Sample size and Bentler and Bonnett's nonnormed fit index. Psychometrika, 51, 375-377.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. Sociological Methods & Research, 17, 303-316.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Jöreskog & H. Wold (Eds.) Systems under indirect observation, part 1. New York: North Holland Publishing Company.

Boomsma, A. (1983). On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality. Unpublished doctoral dissertation, University of Groningen, Groningen.

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

Bray, J. H. & Maxwell, S. E. (1982). Analyzing and interpreting significant MANOVA's. Review of Educational Research, 52, 340-367.

Bray, J. H. & Maxwell, S. E. (1985). Multivariate analysis of variance. Beverly Hills, CA: SAGE Publications.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37, 62-83.

Gallini, J. K. & Mandeville, G. K. (1984). An investigation of the effect of sample size and specification error on the fit of structural equation models. Journal of Experimental Education, 53, 9-19.

Gerbing, D. W. & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. Multivariate Behavioral Research, 20, 255-271.

Geweke, J. F. & Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. Journal of the American Statistical Association, 75, 133-137.

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.

Hoetler, J. W. (1983). The analysis of covariance structures: goodness-of-fit indices. Sociological Methods & Research, 11, 325-344.

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

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

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

IMSL (1982). IMSL Library, Reference Manual. Houston, TX: International Mathematical and Statistical Libraries.

Ito, K. (1969). On the effect of heteroscedasticity and nonnormality upon some multivariate test procedures. Multivariate Analysis, 2, 87-120.

Ito, K. (1980). Robustness of ANOVA and MANOVA test procedures. In P.R. Krishnaiah (Ed.) Handbook of statistics, vol. 1. (pp. 199-235). New York: North Holland Publishing Company.

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

Jöreskog, K.G. & Sörbom, D. (1986) LISREL VI: Analysis of linear structural relationships by maximum likelihood, instrumental variables, and latent squares methods. Mooresville, Indiana: Scientific Software.

Kenny, D. A. (1979). Correlation and causality. New York: John Wiley & Sons.

Kühnel, S. M. (1988). Testing MANOVA designs with LISREL. Sociological Methods & Research, 16, 504-523.

La Du, T. J. & Tanaka, J. S. (1989). Influence of sample size, estimation method, and model specification on goodness-of-fit assessments in structural equation models. Journal of Applied Psychology, 74, 625-635.

Lomax, R. G. (1983). A guide to multiple-sample structural equation modelling. Behavior Research Methods & Instrumentation, 15, 580-584.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: the effect of sample size. Psychological Bulletin, 103, 391-410.

McDonald, R. P. (1978). A simple comprehensive model for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 31, 59-72.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. Psychometrika, 49, 115-132.

Muthén, B. O. (1987). LISCOMP: Analysis of linear structural equations with a comprehensive measurement model [User's Guide]. Mooresville, IN: Scientific Software.

Muthén, B. O. (1989). Latent variable modelling in heterogeneous populations. Psychometrika, 54, 557-585

Rock, D. A., Werts C. E. & Flaugher, R. L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. Multivariate Behavioral Research, 13, 403-418.

Scheffé, H. (1943). On solutions of the Behrens-Fisher problem based on the t-distribution. Annals of Mathematical Statistics, 14, 35-44.

Sharma, S., Durvasula, S., & Dillon, W. R. (1989). Some results on the behavior of alternate covariance structure estimation procedures in the presence of non-normal data. Journal of Marketing Research, XXVI(May), 214-221.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. British Journal of Mathematical and Statistical Psychology, 27, 229-239.

Sörbom, D. (1981). Structural equation models with structured means. In K.G. Jöreskog and H. Wold (Eds.). Systems under indirect observation: causality, structure and prediction. New York: North Holland.

Tanaka, J.S. (1987). "How big is big enough?": sample size and goodness of fit in structural equation models with latent variables. Child Development, 58, 134-146.

Tucker, L. R. & Lewis, C. (1973). The reliability coefficient for maximum likelihood factor analysis. Psychometrika, 38, 1-10.

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