The Report committee for Kevin Richard Higginbotham

Certifies that this is the approved version of the following report

Identification of Variables Contributing to Group Differences in Descriptive Discriminant Analysis

APPROVED BY

SUPERVISING COMMITTEE

Supervisor: ____________________________

Keenan Pituch

____

Tiffany Whittaker

Identification of Variables Contributing to Group Differences in Descriptive Discriminant Analysis

By

Kevin Richard Higginbotham, B.A; M.B.A.

Report

Presented to the Faculty of the Graduate School
of the University of Texas at Austin
in Partial Fulfillment
of the Requirements
for The Degree of
Master of Arts

THE UNIVERSITY OF TEXAS AT AUSTIN
May 2014

Identification of Variables Contributing to Group Differences in Descriptive Discriminant Analysis

By

Kevin Richard Higginbotham, M.A.

The University of Texas at Austin 2014

SUPERVISOR: Keenan Pituch

The identification of predictor variables that meaningfully contribute to group differences in Descriptive Discriminant Analysis (DDA) has had conflicting guidance in the historical quantitative psychological literature. Early simulation results that tested the bias and power of the standardized coefficients and the structural coefficients were ambiguous, yet a consensus still emerged that the structural coefficients were preferred. This study reviews the historical debate and known statistical weaknesses of both standardized coefficients and structure coefficients, summarizes relevant research and proposes a Monte Carlo study that will test whether the inclusion of standardized coefficients in interpreting DDA results for both the two-group and three-group cases can assist applied researchers in meaningfully ranking variables contributing to group differences.

Table of Contents

Introduction

Linear Discriminant Analysis (LDA) is a core multivariate technique used in several lines of academic inquiry that extracts uncorrelated linear combinations from a matrix of 2 or more predictors to describe a known categorical division. Analytically, canonical methods such as LDA subsume the univariate and parametric methods as 'special cases' and provide the protection against Type 1 'experimentwise' error inflation when testing multiple hypotheses [Thompson, 1991]. Descriptive LDA analyzes the categorical divisions for two or more categories, though the fundamental article introducing the technique focused on only two categories [Fisher, 1936]. The extension to 3 or more categories is generally attributed to Rao in 1948. The category for each set of predictors is known a priori by the researcher but any differences in the vector of means between the categories are generally not assumed. This question is formally tested by using multivariate analysis of variance (MANOVA) with the null hypothesis of $H_0 = u_1 = u_2 = \cdots u_j$, where $\acute{u}_j = [u_{1j}, u_{2j}, \dots, u_{pj}]$ and $\acute{u}_j$ represents the vector of outcome variable means (the centroids). The Wilks' Λ (among others) can then be used as an omnibus test to determine if further contrast testing or discriminant analysis is in order [Huberty & Olejnik, 2006]. Formally, the null hypothesis tests the condition of no difference between any of the centroids.

Describing specific group differences after the null hypothesis from MANOVA has been rejected is one of the primary applications of 'descriptive' LDA [Stevens, 2009]. Indeed, Fisher and Rao's articles were concerned with taxonomies of biological and ethnographic distinctions. Another application that has proven increasingly useful is the classification of future cases into one of several groups based on the observed vector's likelihood of occurrence. Predictive LDA, as this application is generally referred to in the literature, has taken on an increasing importance in the past two decades as one of several algorithms found useful in computer applications of unsupervised machine learning [Ripley 1996]. Philosophically there has been some disagreement about which application of LDA, predictive or descriptive, should precede the other (Huberty & Olejnik, 2006), but regardless the techniques remain 'closely aligned.' The outcome of Predictive LDA is typically judged by the technique's ability to maximize classification accuracy for a given data set (defined by its 'hit rate'), while LDA's objectives (the topic of this research paper) are realized by interpreting the resulting discriminant functions to better characterize the distinctions between the groups with regards to the underlying predictor variables. While interpreting the underlying discriminant functions can have value in predictive LDA, it is of paramount importance to a researcher attempting to describe group differences with descriptive LDA– and some disagreement continues in the literature with regards to best practices around this question.

Stevens (2009), advocates two primary methods to interpret linear discriminant functions – the standardized partial regression coefficients (Std b's), defined as the raw coefficient for each predictor multiplied by its standard deviation, and the structural coefficients (structure-Rs), defined as the correlation between the discriminant function(s) and each of the underlying variables. This recommendation sits upon extensive discussion in the literature with regards to the strengths, weaknesses and situational appropriateness for each method – analyses which can differ in the concreteness of their conclusions depending upon whether 2 or 3+ categories are analyzed. Tatsuoka (1973) argued that each approach has its uses as long as 'we keep their different objectives in mind.' Broadly, these differing 'objectives' center around whether the researcher wants to fully consider the multivariate structure when interpreting the relative contribution of the predictor variables, or, rather, emphasize the distinct contribution of each variable – more akin to a univariate interpretation. Structural coefficients are widely acknowledged to ignore the presence of other variables when quantifying the relationship between the predictors and the discriminant function

[Finch 2010]. However, standardized coefficients have been characterized as unstable, as their values will change if certain variables are deleted or added [Rencher, 1992]. Two early Monte Carlo simulation studies in 1975, one performed by Barcikowski and Stevens, the other by Huberty, attempted to bring some clarity to the debate. However, both studies did not find evidence for favoring one coefficient over the other unless sample sizes exceeded 100. Nevertheless, Huberty and Wisenbacker extended the criticism of standardized coefficients in 1992 when they asserted that reliance upon the std-b's results in 'dubious generalizability' given their sampling fluctuation.

Given this shaky historical foundation, many methodologists have still made strong claims regarding the superiority of each coefficient. In 1992, Huberty & Wisenbacker argued that there 'is little doubt' that the most popular approach to ordering outcome variables was to use the standardized coefficients while arguing that the structure-Rs are, in fact, more appropriate. However, Finch in 2010 did not include standardized coefficients in his simulation study to determine meaningful thresholds for interpreting structure-Rs citing the dominant interpretation of most textbooks was to recommend the use of structural coefficients. Yet, major implementations of LDA software have now been designed to provide only the unstandardized coefficients and make no mention of the structure-Rs – leading to further methodological confusion [Venables & Ripley, 2002]. This confusion was furthered by the results of Finch's thorough experimental design in 2010 that showed no compelling recommendations regarding the interpretation of structure-Rs for both standard rules of thumb or bootstrapping at N's below 100. For N's above 100, Finch did find that Bootstrapping was the 'best approach' although power could drop to as low as .85 and the type 1 error rate was still consistently 9% for the 100 and 150 sample size factors. These findings resonate all the more given that Huberty and Wisenbacker in 1992 recommended a more 'heuristic' approach given (at that time) the lack of widespread access to the advanced computing that would allow more sophisticated bootstrap simulations and, implicitly, verify the superiority of the structure-Rs . Now that the computing gap has been closed, Finch's study was able to cross many more factors in exploring the structure-Rs than the older simulation studies, yet the results were still somewhat lackluster. At a minimum, the question remains open as to which coefficient would better assist an applied researcher attempting to interpret a linear discriminant function. Barring a clear answer to the question of which coefficient is superior, additional insight could still be gained by determining how the two coefficients might be used in conjunction with one another to improve Finch's estimates of power and type-1 error in an applied descriptive LDA study.

This study hopes to further assist methodologists by extending Finch's 2010 paper to include standardized coefficients and determine if their performance rivals that or exceeds structural coefficients in interpreting which predictor variables are most contributing to the linear discriminant function. In addition, it will investigate whether a combination of the two coefficients as described by Stevens in 2009 may also further assist in interpretation. Finally, it will extend the experiment to encompass both the two and three category conditions to determine if an increase in the classification dimension alters the effectiveness of using one or both of these coefficients to describe group differences in an applied DDA.

Descriptive Discriminant Analysis and Interpretation

DDA for Two Groups:

The discriminant function for two groups is represented by the linear combination $z = \boldsymbol{a}'\boldsymbol{y}$ in which a row vector of coefficients $\boldsymbol{a}$ is sought such that the distance between the two group mean vectors is maximized [Stevens, 2009]. The technique assumes that the two populations have equal covariance matrices yet different mean vectors u_1 and u_2. Sample observations are distinguished by groups such that $y_{11}, y_{12} \ldots y_{1n_1}$represent measurements on p variables for group 1 and $y_{21}, y_{22} \ldots y_{2n_2}$ represent measurements on the same variables for group 2 with n representing the number of observations in each group. It is a strength in discriminant analysis that the n's are not required to be balanced, yet the number of observations in group 1 ($n1$) and group 2 ($n2$) must be greater than p such that $n1 + n2 - 2 > p$. The single discriminant function for the two group case represents the linear combination that maximizes the distance between the two group mean vectors by 'transforming each observation vector into a scalar' [Rencher 2002]:

$$z_{1i} = \boldsymbol{a}'\boldsymbol{y}_{1i} = a_1 y_{1i1} + a_2 y_{1i2} + \cdots + a_p y_{1ip},\ i = 1, 2, \ldots, n1 \qquad [1]$$

$$z_{2i} = \boldsymbol{a}'\boldsymbol{y}_{2i} = a_1 y_{2i1} + a_2 y_{2i2} + \cdots + a_p y_{2ip},\ i = 1, 2, \ldots, n2 \qquad [2]$$

After each observation vector is compressed into a scalar for each individual the respective means of these scalars can be calculated by $\bar{z}_1 = \sum_{i=1}^{n1} z_{1i}/n1$ and $\bar{z}_2 = \sum_{i=1}^{n2} z_{2i}/n2$ or, equivalently, $z_1 = \boldsymbol{a}'\overline{\boldsymbol{y}}_1$ and $z_2 = \boldsymbol{a}'\overline{\boldsymbol{y}}_2$ where $\overline{\boldsymbol{y}}_1$ and $\overline{\boldsymbol{y}}_2$ represent the vector of means for each variable in groups 1 and 2. The analytical task then becomes finding a single discriminant function $\boldsymbol{a}$ that maximizes the standardized difference between the n x 1 vector of composite means for each group, $\bar{z}_1$ and $\bar{z}_2$, such that:

$$\frac{(\bar{z}_1-\bar{z}_2)^2}{s_z^2} = \frac{[\boldsymbol{a}'(\bar{\boldsymbol{y}}_1-\bar{\boldsymbol{y}}_2)]^2}{\boldsymbol{a}'\boldsymbol{S}_{pl}\boldsymbol{a}} \qquad [3]$$

Note that the maximum of Equation (3), as shown by Rencher [2006], occurs at:

$$\boldsymbol{a} = \boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2) \qquad [4]$$

Or, when $\boldsymbol{a}$ is 'any multiple' of $\boldsymbol{S}_{pl}^{-1}(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)$ with $\boldsymbol{S}_{pl}^{-1}$ representing the inverse of the matrix of pooled variance. As Rencher, Huberty and Ripley all point out, this vector is not unique – only its direction and relative values of the various rotations of $\boldsymbol{a}$ are unique. This difference can also be expressed as a matrix operation by substituting equation (4) into (3). The resulting equation is the product of the transpose of the difference in means, the inverse of the matrix of pooled variance and the vector of mean differences:

$$\frac{(\bar{z}_1-\bar{z}_2)^2}{s_z^2} = (\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2)'\,\boldsymbol{S}_{pl}^{-1}\,(\bar{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}_2) \qquad [5]$$

Thus, the maximum difference between $\boldsymbol{a}'\bar{\boldsymbol{y}}_1$ and $\boldsymbol{a}'\bar{\boldsymbol{y}}_2$ can be achieved as long as the critical assumptions of independence, multivariate normality and equal covariance matrices are met. The result of the discriminant function is a 1 x n matrix for each group representing the composite z-score for each individual in each group.

DDA for 3 or more Groups

Linear Discriminant Analysis for several groups extends the goal of the two-group case and calculates the linear combination of variables that best separates k groups measured by multivariate observations [Huberty, 2006]. While the results of an LDA for several groups cannot be expressed as the singular point of maximal separation as in the two group case, multi-group LDA does provide an opportunity to visualize group separation on three or more observed variables upon a minimum of two discriminant functions [Venables & Ripley, 2002]. Equally important, interpreting the contributions of the underlying variables to each of the discriminant functions can provide insight into the relative contributions of the multivariate predictors to group separation. This notion of 'relative' contribution becomes critical in deciphering the mathematical difference between the standardized coefficients and the structure-Rs. It is at the heart of the scholarly disagreement between Rencher and Huberty and to help underscore this point the initial multi-group discriminant analysis model is presented below.

The description of the multiple group model is as follows: For $\boldsymbol{k}$ groups with n_i observations in the i_{th} group (with $i = 1,2, \dots k$) each observation vector $\boldsymbol{y}_{ij}$ is a 1 x $\boldsymbol{p}$ vector for individual j where $j = 1,2, \dots n_i$ [Rencher, 2002]. This vector is transformed to a scalar z value by the matrix expression $z_{ij} = \boldsymbol{a}'\boldsymbol{y}_{ij}$. Means can then be found by $\bar{z}_i = \boldsymbol{a}'\overline{\boldsymbol{y}}_{\boldsymbol{i}}$ where $\overline{\boldsymbol{y}}_{\boldsymbol{i}} = \sum_{j=1}^{ni} \boldsymbol{y}_{ij}/n_i$. Again, we seek a vector $\boldsymbol{a}$ that provides maximum separation between the computed means of the linear composites for each of the k groups $\overline{z}_1, \overline{z}_2, \dots, \overline{z}_{1k}$. As shown by Rencher in 2002, equation 3 above can be extended to the multi-group case by substituting the H matrix (the between group variation) into the numerator and the E matrix (the within group variation) into the denominator. Both equations are presented below for group ***i*** and individual ***j***:

$$\boldsymbol{H} = \sum_{i=1}^{k} n_i(\overline{\boldsymbol{y}}_{\boldsymbol{i}.} - \overline{\boldsymbol{y}}_{..})(\overline{\boldsymbol{y}}_{\boldsymbol{i}.} - \overline{\boldsymbol{y}}_{..})' = \sum_{i=1}^{k} \frac{1}{n_i} \boldsymbol{y}_{i.}\boldsymbol{y}'_{i.} - \frac{1}{N} \boldsymbol{y}_{..}\boldsymbol{y}'_{..} \qquad [6]$$

$$\boldsymbol{E} = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{k}} \sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n_i}} (\boldsymbol{y}_{\boldsymbol{ij}} - \overline{\boldsymbol{y}}_{i.})(\boldsymbol{y}_{\boldsymbol{ij}} - \overline{\boldsymbol{y}}_{i.})' = \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{k}} \sum_{\boldsymbol{j}=\boldsymbol{1}}^{\boldsymbol{n_i}} \boldsymbol{y}_{\boldsymbol{ij}}\boldsymbol{y}'_{\boldsymbol{ij}} - \sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{k}} \frac{\boldsymbol{1}}{\boldsymbol{n_i}} \boldsymbol{y}_{i.}\boldsymbol{y}'_{i.} \qquad [7]$$

In equations 6 and 7, i represents group i from the sequence of groups 1:k and individual j from the number of individuals in each group sample. Utilizing the H and E matrices in this way allows the differences in group means to be accounted for in all groups and is simplified into matrix notation below:

$$\lambda = \frac{\boldsymbol{a\prime Ha}}{\boldsymbol{a\prime Ea}} \qquad [8]$$

Cross multiplying this can be further simplified to:

$$\boldsymbol{a'Ha} = \lambda \boldsymbol{a'Ea} \qquad [9]$$

Analytically, the value of $\boldsymbol{a}$ is sought such that the maximum of λ is reached with additional solutions derived from $(\boldsymbol{E}^{-1}\boldsymbol{H} - \lambda\boldsymbol{I})\boldsymbol{a} = 0$. In this way, ranked Eigenvalues $\lambda_{1,}\lambda_2, \ \dots \ \lambda_s$ and their related eigenvectors $a_1, a_{2\dots}a_s$ are created such that $\lambda_1 > \lambda_2$ and $\lambda_2 > \lambda_3$, etc. The number of eigenvalues extracted is equivalent to the rank of H, which will be the smaller of k-1 or p [Huberty, 2006]. From the number of extracted eigenvectors, s, a corresponding number of discriminant functions are created such that the eigenvalues (z_1, z_2, z_s) are created by the following discriminant functions - $z_1 = \boldsymbol{a'}_1\boldsymbol{y}, \ z_2 = \boldsymbol{a'}_2\boldsymbol{y} \ \dots \ z_s = \boldsymbol{a'}_s\boldsymbol{y}$. These functions are uncorrelated but are not orthogonal as $\boldsymbol{a'}_i\boldsymbol{a}_j = 0$ for $i \neq j$ as $\boldsymbol{E}^{-1}\boldsymbol{H}$ is not symmetric (Rencher, 1998). Finally, since these functions are uncorrelated it is a straightforward calculation to determine the relative importance of each eigenvalue's contribution towards the maximization of group differences:

$$\frac{\lambda_1}{\sum_{j=1}^{s} \lambda_j} \qquad [10]$$

So, at this stage the analyst will have either $k - 1$ or p discriminant functions and will also have an initial metric of the importance of each. These functions can be tested for significance by using the Wilks Λ-test first on all of the eigenvalues such that $\Lambda_1 = \prod_{i=1}^{s} \frac{1}{1+\lambda_i}$ and, if significant, the largest of the eigenvalues can be assumed to be significant as well [Stevens, 2009]. The remaining eigenvalues can then be iteratively tested using the same approach. It is generally acknowledged in the literature that functions associated with small eigenvalues can be neglected and often 'two or three' functions will suffice to describe the separation. Where there is less agreement in the historical literature is the relative importance of the underlying predictor variables in contributing to the discriminant functions. So, it is to these metrics that we will turn our attention to next.

Standardized Discriminant Function Coefficients

The construction of the $\boldsymbol{a}$ vector containing the coefficients of a particular linear discriminant function clearly possesses intuitive strengths in assessing each predictor $y's$ contribution to group separation. However, there is an additional advantage to standardizing the $\boldsymbol{a}$ vector so that each individual $y's$ contribution can be adjusted to the same scale given that many collected datasets will not have inherently comparable variances [Rencher, 2002]. For the two group case we can express the discriminant function in terms of standardized variables as shown below:

$$z_{1i} = a_1^* \frac{y_{1i1} - \bar{y}_{11}}{s_1} + a_2^* \frac{y_{1i2} - \bar{y}_{12}}{s_2} + \cdots + a_p^* \frac{y_{1ip} - \bar{y}_{1p}}{s_p}, i = 1,2, \ldots n_1 \qquad [11]$$

$$z_{2i} = a_1^* \frac{y_{2i1} - \bar{y}_{21}}{s_1} + a_2^* \frac{y_{2i2} - \bar{y}_{22}}{s_2} + \cdots + a_p^* \frac{y_{2ip} - \bar{y}_{2p}}{s_p}, i = 1,2, \ldots n_2 \qquad [12]$$

The composite z_{1i}variable is computed by dividing each observation's difference from the within group mean by the within sample standard deviation s_r (for the r_{th} variable). The two-group standardized case can be further simplified to matrix notation by taking the square root of the r_{th} diagonal of our S_{pl} matrix and multiplying it by our initial vector of linear discriminants, $\boldsymbol{a}$**.**

$$\boldsymbol{a}^* = (diag\ \boldsymbol{S}_{pl})^{\frac{1}{2}}\ \boldsymbol{a} \qquad [13]$$

And, intuitively, this standardizing function can be extended to the multi-group case by denoting the r_{th} coefficient in the m_{th} discriminant function as $a_{mr}, m = 1,2, \ldots, s; r = 1,2, \ldots, p$ creating the standardized form of the within group standard deviation (pulled from $\frac{\boldsymbol{E}}{v_E}$ for the multi-group case) multiplied by the several discriminant functions that will exist in the multi-group case.

$$a_{mr}^{*} = s_r a_{mr} \qquad [14]$$

The resulting variables are now 'scale-free' and, in Rencher's words 'correctly reflect the joint contribution of the variables to the discriminant function z (as presented for the two-group case in equations 11 and 12) that maximally separates the groups [2002]. This property also extends to the correlations among the matrix variables in the multi-group case as each discriminant function's coefficient vector $\boldsymbol{a}$ when expressed as an eigenvector $E^{-1}H$. Of course, the question still remains on how to interpret the resulting standardized coefficients.

For the two-group case an interpretation routine frequently mentioned in the literature entails examining the standardized coefficients of the resulting discriminant function and ranking the absolute value of the coefficients to determine which underlying variables are contributing most to the group differences [Rencher & Scott, 1990]. For the multi-group case the recommended analytic routine is similar and takes into account the presence of multiple discriminant functions. Since the multiple functions are uncorrelated each of the functions can have its own unique interpretation (tempered by the percentage of variance described by each function) with increasingly narrow bands of separation explained by the succeeding functions. Further insight can be gained by taking into account the signs of the variables. Two well documented limitations of the standardized coefficients revolve around typical shortcomings of linear combinations: mutability in the face of additional variables and stability with regards to sample size [Huberty 2006]. Both will be addressed in detail after the introduction of structure-Rs.

Structural Coefficients

Structural Coefficients represent the correlation(s) between the linear discriminant function(s) and each of the outcome variables. As Finch recounted in 2010, there are two main approaches to calculating these structure-Rs: the first taking into account the total group correlations and the second taking into account only the within group correlations (or, R_w). The total group correlations ($SC_T = R_T D$) have been shown to ignore the differences in group means, a drawback that has generally been agreed upon in the literature to be a significant one (Huberty & Olejnik, 2006), while the correlations to the within matrix ($SC_w = R_w D$) 'corrects' for this shortcoming. Alternatively, as shown by Rencher in 1992, the correlation (or factor loading) for the two group case can be expressed as follows:

$$r_{y_i x} = \frac{\bar{y}_{1i} - \bar{y}_{2i}}{\sqrt{diag(S_{pl})}D^2} \qquad [15]$$

Rencher's analysis shows that D^2 is equivalent to equation 5, and that equation 15 above is proportional to the univariate t-statistic and that use of this resulting coefficient 'unintentionally' reduces the multivariate setting to a univariate one as in equation [16] below:

$$t_i = \frac{\bar{y}_{1i} - \bar{y}_{2i}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) diag S_{pl}}} \qquad [16]$$

The interpretation of these structure-Rs follows the same recommended procedure as the standardized coefficients – assess the absolute value of the magnitude of the association and take the arbitrary signs into account when 'labeling' the LDF. However, additional distinctions are critical when discerning between the two-group case and the multi-group case as described by Huberty and Weisenbacker in 1992. These structure-Rs were proposed by Huberty and Wisenbacker as a way to 'account for the sampling fluctuation of [$\boldsymbol{a}$]' in the two-group case while acknowledging that the multiple-group case was 'more complicated.' While not overtly recommending the structure-Rs in the multi-group case, the lack of generalizability for the standardized coefficients was again noted – and this debate about ranking a variable's contribution to the LDF has even deeper roots in the literature.

Variable Ranking In DDA: Literature Review

The methodological literature regarding the interpretation of group differences from descriptive LDF coefficients has focused on the ranking of such coefficients to assess which of the underlying group of variables contributes most to those differences. It remains to be seen whether or not framing the question in this way has contributed to the ensuing confusion on whether the structure-Rs or standardized coefficients provide more reliable results – that is, the notion of ranking in the multivariate context may itself be problematic [Thompson 1991]. Regardless, the results of most previous simulation studies that attempted to determine the superiority of one coefficient over the other were murky at best.

Two initial efforts were undertaken in 1975, one by Huberty and the other by Barcikowski and Stevens. Both had ambiguous results. Huberty's study featured two group sizes ($k = 3$ & $k = 5$) and compared the performance of standardized coefficients (equation 14) to the structure-Rs (both the total and within correlations) on 10 simulated normal predictors. Huberty's multi-group simulation specified a known common covariance matrix and 'standard normal' scores were created in his simulated sampling. His goal was to create a covariance matrix that would be common to applied researchers which he describes only as 'positive and modest' [Huberty 1975]. The common population matrix (Σ) was constructed according to the classical factor analysis model below (utilizing Huberty's notations when possible):

$$\Sigma = A_{pop} A'_{pop} + D^2_{pop} \qquad [17]$$

A_{pop} is a $10\ x\ m$ matrix of coefficients created from m common factor loadings and D_{pop} is a 10 x 10 diagonal matrix of coefficients of unique factors. Effect sizes were constructed in another $10\ x\ k$ matrix and separation among the groups was achieved by multiplying the common covariance matrix above by the effect size matrix and obtaining the population mean matrix.

Sample sizes were set to 90, 150, 300 and 450 and only analyzed each predictor's contribution to the first LDF. Sample score matrices were generated by $X_g = A\hat{F} + D\widehat{U} + M_g$ in which X_g is the matrix of the observed scores for group g and the $m \times g$ matrix $\hat{F}$ applies the factor loadings to each groups' 10 variables. Huberty recognized that the 'weights in the two models are not directly related except in the two group case, but argued that this approach was valid nonetheless.

The results led Huberty to conclude that 'an index of absolute contribution is. . . out of the question' given that the studied variables move in concert. The only achievable goal given the computing limitations in 1975 was that of determining the relative contribution of each variable. Counts of the rank obtained by each of the 10 variables were tabulated an analyzed for consistency across replications and then tested by evaluating the λ^2 of Kendall's W (a measure of concordance, or agreement among the rankings). His results with regards to study reliability were mixed. Notably, the standardized coefficients did show one major advantage: they were much more accurate in identifying variables that had no contribution to group separation and this advantage may have been underemphasized [Huberty,1975]. An additional Monte Carlo study was run that same year by Barcikowski & Stevens which focused on establishing the reliability of the structure-Rs . While, the study found that the canonical correlations were 'very stable upon replication' the study also concluded that 'there is no solid evidence for concluding that the components are superior to the coefficients' [Barcikowski & Stevens, 1975]. It should be noted that both studies emphasized the instability of each coefficient when sample sizes are small with the number of subjects per variable suggested to be above 40.

Thompson (1991) conducted another more general study comparing structural coefficients to standardized coefficients in the more general case of canonical analysis (Huberty's study was limited to DA which, quoting Thompson, 'is equivalent to a canonical analysis in which group membership is dummy coded'). He cited several researchers from the mid-70's that insisted on the primacy of the structure-Rs – with at least some of this emphasis deriving from a psychometric position rather than a methodological one. Essentially, some early researchers insisted that canonical correlation analysis demands an emphasis on structural coefficients given the 'synthetic' nature of the underlying latent variables. Several early textbooks also emphasized that structure-Rs will be less influenced by sampling error. It was this second point that was most directly addressed by Thompson's study which reinforced the results of Huberty's and Barcikowsi & Stevens' early work: the results of Thompson's experiment suggested that the structure-Rs and standardized coefficients are not differentially sensitive to sampling error [Thompson, 1991]. Both Thompson's and Huberty's studies were 'synthetic' in that they relied on population data that were created to 'reflect desired variations.'

Thompson's method was to simulate multivariate normal populations and randomly sample from that generated data 1,000 times. 64 conditions were tested: number of variables (12, 8, 6 or 10), sample size (3, 10, 25 or 40 observations per variable) and 4 different covariance matrices (all correlations set to zero, within correlations set to 0 and heterogeneous between correlations, between correlations set to 0 and heterogeneous within correlations and a 'homogenous' scenario in which all correlations were the same and above zero. A distinct population was created for all 4 correlation-matrix scenarios with the first zero correlation condition representing the null hypothesis. Thompson's study allowed for the direct comparison of the standardized coefficients and structure-Rs for the same variable set – in particular, he focused on the 'mean of the mean deviations in a given matrix from the known true population parameters' Thompson [1991]. He found the mean matrix deviations for the populations were consistently small and homogenous and provided a good basis for comparison. The average means of the mean matrix deviations from the 64 parameters were comparable for the standardized coefficients (-.035) and structure-Rs (-.034) for the null hypothesis case and (-.042) and (-.044) for the heterogeneous/between scenario. Standard deviations for the mean matrix deviations were actually somewhat tighter for the standardized coefficients – but Thompson hesitated to generalize this finding.

Two articles published in 1992 further clarified the opposing viewpoints. In the first, Rencher [1992] analytically addressed the conceptual or psychometric issue by first demonstrating that in the two-group case the structure-Rs are equivalent to a univariate t-test on each of the observed variables (see equations 15 and 16). His earlier work in 1988 had already constructed a similar proof with regards to the multi-group case. When comparing k groups on the predictor vector y_i, Rencher showed that the F-statistic is simply a function of the correlation between y_i and all of the extracted structural coefficients [Rencher, 1988]:

$$\sum_{j=1}^{k} r_{x_i y_1}^2 \lambda_j = \frac{b_{ii}}{w_{ii}} = \frac{k-1}{n-k} F_i \qquad [18]$$

Again, Rencher's point, analytically derived, was that the correlations between the variables (x_i) and the canonical discriminant functions j, with the eigenvalues λ_j representing the diagonal elements of B (b_{ii}) and W (w_{ii}), provide 'no information about the multivariate contribution of a variable.' The influence of the presence of other variables is lost, so the alleged psychometric clarity of the structure-Rs cited by previous researchers is akin to reducing a multivariate problem to a univariate one. For this reason, Rencher advocated the use of standardized coefficients [Rencher, 1992]. He argued, with some force, that there was no 'middle-ground' between the univariate and multivariate approaches and the implication was that by seeking additional clarity interpretive errors will be made by applied researchers if they insist on the static structure-Rs when interpreting LDFs.

Huberty and Wisenbacker responded in 1992 by framing the question somewhat differently in recommending that the standardized coefficients be altered to account for their sampling fluctuation. They begin their paper by emphasizing that the issue of determining rank or variable importance can be difficult in that 'statisticians rarely concern themselves with the problem of rank'

- implying that the issue of ranking variables may be an imposition of theory rather than computation. While acknowledging Rencher's analytical proof that the structure-Rs do not take into account the interconnectedness of the multivariate predictor matrix, Huberty & Wisenbacker nevertheless maintain that the use of standardized coefficients still need to be tempered due to problems of sampling fluctuation. This claim is applied to the two-group and multi-group case, with the critical distinction that the two group case can be accounted for by taking $\mathrm{a}_{\mathrm{i}}^{*}/\mathrm{r}^{\mathrm{ii}}$ and acknowledging that the multi-group case is 'more complicated' [Huberty & Wisenbacker, 1992]. The paper goes onto suggest two methods focusing on the 'F-to-Remove' values in order to determine a ranking system that does not over-react to small differences in the F-values for the multi-group case.

Thus, the dismissal of standardized coefficients due to their sampling fluctuation would seem to have its roots in an important paper that does not have the benefit of an empirical test. More puzzling is that Thompson's 1991 study (among others) did provide some evidence that the standardized coefficients were not any more susceptible to sampling error than the structure-Rs. Nevertheless, Huberty & Wisenbacker's criticism morphed into a full-blown dismissal of standardized coefficients in Huberty & Olejnik's 2006 text 'Applied MANOVA and Discriminant Analysis' in which reference to the 1992 article as well as an additional text published by Joy and Tollefson are cited as reasons for which standardized coefficients are not suitable for variable ranking. It should be noted that Joy and Tollefson's article in the Journal for Financial and Quantitative Analysis recommended against standardized coefficients but did not recommend that structure-Rs be used – rather a competing coefficient that calculated the "the portion of the discriminant score separation between the groups, $(\bar{z}_1 - \bar{z}_2)$, that is attributable to the j_{th} variable" [Joy & Tollfeson, 1975] . Fisher later demonstrated that these two methods were commensurate, and that the differences in interpretation were due to the violation of the underlying LDA assumptions (in a cited applied finance case two key variables exhibited significant collinearity) [Fisher, 1978]. Finally, Huberty and Wisenbacker did not clarify how structure-Rs might remedy the sampling problem in the two-group case other than by implying that the correlations may do so.

Rencher responded to these criticisms by acknowledging that the standardized discriminant functions are indeed subject to the traditional limitations of linear combinations used in a regression equation: 1) that the coefficients may well change if other predictors are added or deleted and 2) sample size must be adequate when compared to the number of predictors in order to be generalizable (Rencher, 2002). However, he is also careful to emphasize that with regards to the first property, in a multivariate context 'this is exactly how we want them to behave.' That is, in any multivariate context a researcher must, by definition, be interested in how the variables move in concert. The focus of the disagreement then would logically pivot upon the question of sample size. That is, given the agreement on both sides of the debate that the structure-Rs do not take into account the presence of the other predictor variables (and that the standardized coefficients do), the recommendation to discard standardized coefficients must be due to an increased susceptibility to sampling fluctuation at insufficient N's. One (of many) key questions becomes, do structure-Rs perform better when sample sizes are small? If so, there may be a reason social researchers should prefer them, as smaller sample sizes are more prevalent in psychological research. But, a close reading of the literature has shown that there is no empirical evidence that this is the case and this lack of evidence in the historical literature may well be due to the previous inability to estimate the statistical significance of each coefficient in LDA. Of course, we now know bootstrapping can be used for just such a purpose.

The lack of a reliable method to estimate the statistical significance of structure-Rs and standardized coefficients may have contributed to the unnecessary dismissal of standardized coefficients in the literature. Resampling methods offer a way to provide statistical tests when these sampling distributions are not easily specified [Dalgleish, 1994]. Dalgleish used both the jackknife method and the bootstrap method to assess the statistical significance of the structure-Rs but purposefully omitted the standardized coefficients citing the 'substantive' superiority of the structure-Rs. Again, his implication was that the psychological literature prefers the interpretive superiority of the structure-Rs.

The jackknife technique constructs initial sample statistics using all of the data and then divides the existing data set up into 'slightly reduced bodies of data' (usually one or two), then omits the reduced data set and calculates pseudovalues designated below as $\hat{\theta}_j^*$. The variable k represents the number of datasets the original dataset is divided into and θ_F is the coefficient of interest:

$$\hat{\theta}_j^* = k * \hat{\theta}_F - (k-1) * \hat{\theta}_{-j} \qquad [19]$$

The mean and standard error of the k pseudovalues can be constructed and differences from zero tested by dividing the jackknife estimate by the standard error.

The bootstrap technique treats the sample as a population and then resamples of size n (with replacement) are extracted from the original sample and the statistics of interest (in Dalgleish's study, only the structure-Rs) can be computed. Repeating this resampling multiple times allows the analyst to create a mean bootstrap estimate of the coefficient $\hat{\theta}_B$ and its standard deviation $\hat{\sigma}_B$ with an implied normal distribution. A more formal hypothesis tests can be constructed (with θ as the value of the targeted coefficient under the null hypothesis) according to the formula below:

$$Z = \frac{\hat{\theta}_B - \theta}{\hat{\sigma}_B} \qquad [20]$$

One particular problem in resampling the results of a DDA is the susceptibility of both DDA coefficients to permutations and/or changes in sign similar to that highlighted by Clarkson in his 1979 study that extracted jackknifed estimates of rotated factor loadings. For DDA, with 100's or 1000's of bootstrap samples there will be cases in which the sign or function of a standardized coefficient or structure-R may change from one sample to the next. Both Finch [2010] and Dalgleish [1994] followed Clarkson's approach and changed the signs of the structure-Rs and the function order so as to minimize the sum of squared differences between the full sample and the bootstrapped resamples. Dalgleish's study highlights other complexities in estimating the statistical significance of the interpretive coefficients in DDA. He first collected data from a 1980 study that examined the interaction between particular types of crime and personality traits. A full DDA analysis was run on a small subset of this field collected data and estimates of the structural coefficients were obtained and the significance of two discriminant functions confirmed by testing with the Wilk's Criteria. Bootstrapped and jackknifed estimates of the mean were then obtained and standard errors for the population were estimated. Bootstrapped estimates were created with 1000 resamples while the jackknife left out one observation at a time. The results showed several significant discrepancies between the 95% bootstrapped confidence intervals and jackknifed estimates as well as conflicting interpretations on whether to include some of the weaker associated structure-Rs. It was also true that standard errors for the second LDF showed more variability than the first.

Given this interpretive quandary, Dalgleish [1994] ran a Monte Carlo study to determine if 'the jackknife tests and the bias corrected bootstrap are liberal or that the standard and percentile bootstrap are conservative.' This required that the difficult task of simulating data with known structure-Rs be completed which Dalgleish tackled by creating a SAS macro that generated data 'based on the structural equation formulation of canonical correlation analysis.' The problem of alignment reemerges with this approach. Given that structure-Rs will not necessarily align with the population values Dalgleish simply discarded all non-aligning data sets and kept only those data sets that did align. The data was created such that standard-Rs below the threshold of .15 were simulated to be 0 and only those that were found to be above the nominal cutoff of .3 were simulated to have an effect. Thus, type-1 error and coverage could be assessed for both jackknife techniques and the three bootstrapped confidence intervals: the standard bootstrap of 95%, the percentile bootstrap of 2.5% & 97.5% and the bias corrected bootstrap [Effron & Gong , 1983]. With regards to type-1 error, the results showed the jackknife methods 'too liberal' with estimated coefficients roughly 1.5 to 2x the nominal α levels of .01, .05 and .10. The bootstrap method performed 'very well' if not conservatively for these same α levels. Coverage was also assessed for the 100 data sets and again, both jackknife methods performed poorly, as well as the bias-corrected bootstrap. It should be noted that the strongest technique with regards to coverage (the standard bootstrap) had a 95% confidence interval that captured the largest structure-R of .75 only 88% of the time while weaker structure-Rs were captured as high as 97% – an unintuitive finding that may reinforce Rencher's analytical criticisms of the structure-Rs.

In 2010 Finch revisited the question of which method of interpreting the structure-Rs is superior. Finch's well researched work revisited many of the central questions of rank-ordering predictors for their contribution to group separation including providing an experimental distinction between the structure-Rs when computed with total group correlations (RC_T) -- which do not account for group mean differences -- and within-group correlations (RC_W) which do account for such differences. Interestingly, Finch advanced the historical debate over DDA interpretation from that of simply rank-ordering the predictors to also establishing a threshold for when a predictor meaningfully contributes to group differences. Citing two major quantitative psychology texts (Pedhazur 1997 & Tabachnick & Fidell 2001) Finch proposed to test established 'cut' points by which the contribution of a particular variable could be deemed significant (both texts suggest an SC value of .3 as being important). Additionally, relative value ranking and Dalgleish's standard bootstrap approach were rolled into his study. He classified these broadly as 'cutoff' methods, relative ranking (such as Huberty & Olejnik and Stevens) and the inferential approach (the standard bootstrap).

Taking his cue from Huberty & Olejnik, Finch did not include the standardized coefficients in his study. This omission was not haphazard and was well-considered given the established literature. First and foremost, Finch correctly observed that the early simulation studies to determine the superiority of one coefficient over the other were not conclusive and given this finding a researcher is reasonable in selecting just one. The strong bias in quantitative psychology texts for the structure-Rs gave Finch the historical justification to focus solely on these to the exclusion of the standardized coefficients. Additionally, Finch correctly pointed out the theoretical shortcomings of each coefficient (mutability for the standardized coefficients and predictor isolation for the structure Rs). However, as indicated in the literature review above, some subtlety to the previous debates may have been overlooked – an issue that will be discussed further in the proposed study for this paper.

Finch's study was also quite thorough but did limit itself solely to the two-group DDA condition with two and six predictors. The interpretation methods were .3, .4 and .5 for the cutoff values as well as the bootstrapped confidence interval. His design also allowed for the ranking of the simulated predictors. Normal and non-normal distributions were also tested given the historical importance of multivariate normality in both DDA and predictive discriminant analysis [Finch, 2010]. Group separation was simulated through the use of Cohen's d (the difference between 2 means divided by the pooled standard deviation (s) for the data):

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s} \qquad [21]$$

$$s = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \qquad [22]$$

Effect sizes of 0, .5 and .8, or ‘no difference, moderate difference and large difference’ with respect to behavioral research were used. Simulated sample sizes were also modest: 30, 60, 100 and 150 and were also varied by the relative size of the groups. Two conditions were tested by Finch – the first in which group sizes were equal and the second in which one group was twice as large. The equality of covariance assumption was also tested with an additional factor of equal covariance matrices and one in which one group had a standard deviation that was five times the other. As mentioned above, both total-group and within-group structure-Rs were tested. Finally, a full factorial analysis of variance was run by Finch on the results to determine which of the manipulated factors was influencing the rate at which the structure-Rs were detecting variable(s) most associated with group differentiation (Finch, 2010).

Finch [2010] defined ‘power’ as the ability to ‘correctly identify variables associated with group separation.’ Type 1 error rate was defined as the inclusion of a non-correlated variable as contributing to group separation. Finch claims that his results strongly recommend an additional interpretive criterion such as the standardized coefficients to assist in DDA interpretation. Direct comparison of structural coefficients had grossly inadequate power in both the normal and non-normal conditions. For example, with an effect size of .5 and 0 in the two variable case, direct comparison was able to identify the significant variable only 58% of the time. Power for the .3 cutoff was highest, but this would be expected given its comparatively low threshold. The standard bootstrap method had power that was comparable to the .5 cutoff value. Type 1 error rate was high for all methods, but was least for the standard bootstrap. Finch’s study found that Type 1 error was ‘high’ in that it was over .1 in all cases and significantly higher than .1 for all methods other than the bootstrap. Power was also slightly lower in the six variable case but for the most part Finch concluded that the results generalize. Finally, it should be noted that standardized coefficients have been shown to correctly identify variables with no effect – a weakness in the structure-Rs that was pronounced in Finch's study presenting an opportunity for further research as his study did not include such coefficients. Clearly, Finch's overall conclusions were stark: 'using only one index, such as the SC's, oversimplifies the process and may not be advisable.'

Finch and Laking ran an additional study investigating standardized coefficients in 2008 (while seemingly published before his 2010 article, the 2008 article was written after his initial paper published in 2010 was presented at a conference in 2007). Their results bolstered Rencher's position that the standardized coefficients better transmit multivariate effects than the structure-Rs in some two-group conditions. For predictor variables that were normally distributed and shared equal covariance matrices, Finch and Laking found the standardized coefficients to be highly accurate in identifying variables that did not contribute to group differences. The coefficients were accurate over 90% of the time in most cases and often exceeded 95% depending on the interaction between the experimental conditions of number of predictors and sample size.

Additionally, when the six simulated variables all had distinct effect sizes (again, generated by Cohen's d), the standardized coefficients correctly ranked the variables between 70% and 95% of the time depending on the simulated correlation between predictors and sample size. For the larger sample size of 150, Finch and Laking found that for normally distributed predictors with homogenous covariance matrices, predictors simulated to differ by .5 in their weights were correctly ordered over 80% of the time, and when all 6 predictors differed by .8 they were correctly ordered over 90% of the time [Finch & Laking, 2008]. Where the standardized coefficients failed miserably was in the case in which one variable was associated with the group difference and the other five were not – a situation that would theoretically call for a univariate analysis – highlighting the importance of an applied researcher correctly identifying the appropriateness of the multivariate approach for a specific research problem. So, Rencher's admonition to not confuse the two worlds of multivariate and univariate analyses rings true in Finch's findings. What the 2008 study lacked was a direct comparison of the standardized coefficients to the structure-Rs and an extension of the findings to the three group condition.

A close reading of the results of previous studies suggests that no experimental evidence has been found that reinforces the often made assumption in the literature that structure-Rs are less susceptible to sampling variation. Thompson commented upon this directly in 1991 when he argued that his study and the previous studies of Huberty and Barcikowski & Stevens 'do not suggest that either structure or function coefficients are inherently sensitive to sampling error.' In the two-group case, Finch clearly demonstrated that the Type 1 error rate of the structural coefficients was unacceptably high [Finch 2010]. His updated study with Laking reinforced this finding – standardized coefficients more accurately identified variables with no contribution to the discriminant function in the presence of multivariate effects [Finch, Laking 2008]. Given these findings a study that directly compares the standardized coefficients with the structure-Rs in the two and three group case should greatly assist the interpretation of group differences in a descriptive DDA as well as extend Finch's specific findings to the three-group case.

Statement of Problem

Increasingly, the experimental evidence has suggested that DDA's sample size requirements as documented by Stevens in 2009 may have distorted many methodologist's views of the standardized coefficients. That is, one clear finding from the previous simulation studies is that small $n's$ can significantly distort the findings for both coefficients yet the standardized coefficients appear to have taken the brunt of the negative assessment due to this universal characteristic. The Barcikowski and Stevens study recommended 40 or more observations per predictor variable in order to ensure a reliable Descriptive Discriminant Analysis study. Finch's study significantly reinforced this finding by concluding that 'all of the methods examined . . . had greater power for larger sample sizes' [Finch, 2010], yet most of Finch's simulated N's fell below the threshold of 40 observations per variable in his 6 variable scenarios for both studies. Stevens recommended 20 in his 2009 text and also recommended that both standardized coefficients and structure-Rs be used, however suggesting that the structure-Rs be used for interpretation and the standardized coefficients be used to determine if a variable is redundant. Regardless, it if can be confirmed that the standardized coefficients are no more susceptible to sampling fluctuation than the structure-Rs then a key objection to standardized coefficients will be significantly muted. This finding would then allow the standardized coefficient's clearly established analytical strengths in a multivariate setting to be emphasized. Dalgleish's 1994 study was the first to employ the bootstrap and jackknife resampling methods to statistically test the structure-Rs, yet his study dismissed standardized coefficients on psychometric grounds. This dismissal carried over into Finch's 2008 work, but Finch's results led him to question whether the structure-Rs alone are sufficient for interpretation. An experiment that extends the resampling approach of Dalgleish to the standardized coefficients should bring substantial insight to the debate and is merited by the existing literature.

Historical criticism of standardized coefficients focused on their lack of generalizability and psychometric inadequacy. Refuting the argument for psychometric superiority would be complex and is beyond the scope of a methodologist. However, the claim that standardized coefficients are more susceptible to sampling error can be analyzed. Structural coefficients should 'generalize' better than the standardized coefficients, an implication which has not been supported by the statistical literature. Furthermore, a proven strength of standardized coefficients – their ability to detect predictors that are not associated with group differences has been suggested in the findings of several simulation studies yet underemphasized in the literature [Huberty 1975, Finch 2010]. More recently, Finch and Laking showed that there was considerable power in standardized coefficients with regards to ordering variables in a multivariate setting as long as linear assumptions were met. And, while Rencher's criticism of the structure-Rs' predictor isolation have been absorbed and relayed in the modern literature, it has not yet been definitively shown in a simulation study that standardized coefficients either 1) perform better or 2) are needed to augment the interpretation of structure-Rs. This study will conduct two experiments to directly compare these two coefficients in both the two and three group conditions. It will build itself upon the simulation work of Huberty [1975], Barcikowski & Stevens [1975], Thompson [1991], Finch [2010] and Finch and Laking [2008].

Research Study

The literature review above has suggested that standardized coefficients may have been prematurely dismissed in the quantitative psychology literature with regards to the ranking and interpretation of variables contributing to group separation in a significant discriminant function. To determine the extent to which standardized coefficients might assist an applied researcher, two separate resampling studies are being proposed that will directly compare the std-b's performance to the structure-Rs. The first study will extend the work of Finch's 2010 two-group DDA study which compared three methods of interpreting structure-Rs: traditional cutoff methods, comparison of relative magnitudes and bootstrapping (due to the poor performance of the jackknife technique in Finch's study, this resampling method is not included in the current proposal). This study will build upon Finch's evaluation of the structure-Rs by including an additional calculation for the standardized coefficients. Both coefficients can then be examined in relation to both 'Power' and 'Type-1 Error' as defined by Finch in his 2010 study (quoted to highlight Finch's acknowledgement that for the proposed testing of the cutoff values there will be no formal hypothesis test per se) . The proposed methods and factors will closely mirror Finch's design and will rely on Cohen's D to simulate group separation.

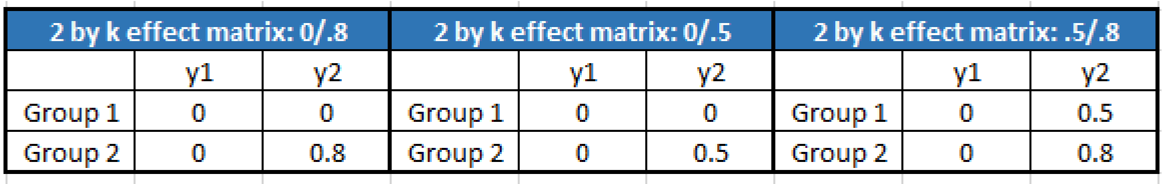
The second study will tackle the more complicated problem of the three-group case. For this simulation study the method will model itself after Thompson's 1994 study and Huberty's in 1975 by specifying uniquely interesting covariance matrices to generate data rather than relying on the univariate metric of Cohen's D in the two-group case. However, I will extend these studies by bootstrapping both the standardized coefficients and structure-Rs to examine Power and Type-1 Error in the more complex multivariate case of interpreting group differences for three-group DDA.

The number of factors will differ from the two group simulation in that the simulation method becomes more complex as will be elaborated upon further below.

Conditions for the Study Comparing Contributions to Group Differences in Two Groups

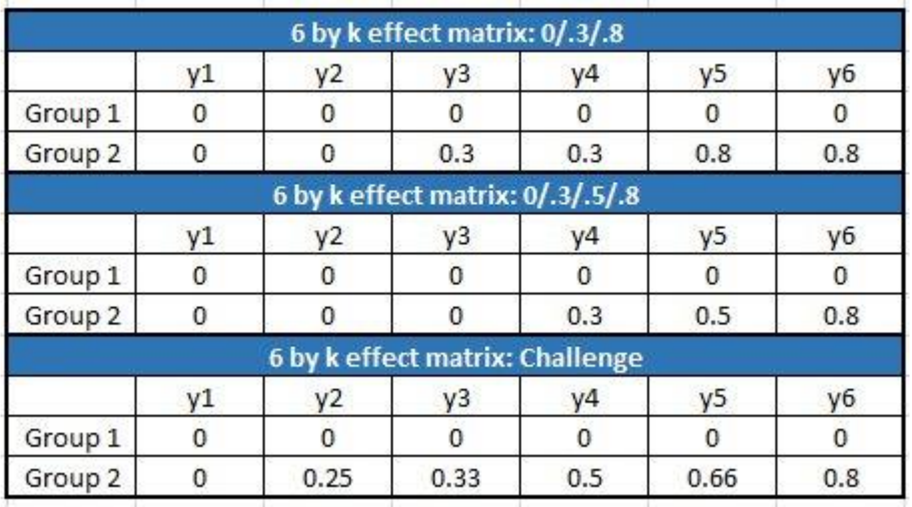
The first proposed study will operate under the two-group condition. The R function lda() authored by B.D. Ripley as part of the MASS package will be used to generate unstandardized DDA coefficients according to equation (4). These scaling coefficients will be transformed first into the standardized coefficients by equation [14] and subsequently into structure-Rs by equation [15]. 1000 Bootstrapped estimates of each coefficient will be generated using the standard Z score [equation 20]. Several conditions shown to impact the performance of both predictive LDA and descriptive LDA will then be varied systematically.

The first condition varied is effect size and will be consistent with those used in the Finch's 2008 study and be limited to sizes of 0, .5 and .8 with the addition of .3 for the 6 variable case. The reasoning is two-fold. First, as noted by Finch, these effect sizes were associated by Cohen with no difference, moderate difference and large difference [Cohen, 1988]. Second, keeping to these values will allow a direct comparison to Finch's initial study. Following Finch's conceptual approach, the rnorm() function in the R software package will be used to generate a 'control' data set for the first group which will be composed of simulated variables of mean 0 and a standard deviation of 1. Group separation will then be created by utilizing Cohen's D (equations 21 and 22) though operationally this will again mirror Finch by plugging the desired effect value of Cohen's d (0, .5 or .8) into the mean parameter of the rnorm() function to achieve the targeted separation. However, the combination of effects will be varied somewhat and a small effect size of .3 will be added to the six variable case so that 1) the task of ranking can be made more clear-cut and 2) conditions that contain multiple variables with no effect in the six-variable case can be constructed – a difference from Finch's design. The blend of effect sizes for the two variable case is shown in the table below:



2 by k effect matrix: 0/.8			2 by k effect matrix: 0/.5			2 by k effect matrix: .5/.8		
	y1	y2		y1	y2		y1	y2
Group 1	0	0	Group 1	0	0	Group 1	0	0.5
Group 2	0	0.8	Group 2	0	0.5	Group 2	0	0.8

And for the six variable case:



6 by k effect matrix: 0/.3/.8						
	y1	y2	y3	y4	y5	y6
Group 1	0	0	0	0	0	0
Group 2	0	0	0.3	0.3	0.8	0.8
6 by k effect matrix: 0/.3/.5/.8						
	y1	y2	y3	y4	y5	y6
Group 1	0	0	0	0	0	0
Group 2	0	0	0	0.3	0.5	0.8
6 by k effect matrix: Challenge						
	y1	y2	y3	y4	y5	y6
Group 1	0	0	0	0	0	0
Group 2	0	0.25	0.33	0.5	0.66	0.8

Unlike Finch's study, I will leave two and three unrelated variables in the second group and only utilize 5 of the 6 variables once (the 'Challenge' condition). This will be to explicitly test each coefficient's ability to correctly detect multiple variables that are not contributing to group effects in the first two 6 variable conditions. As the effect conditions will not be fully crossed, this factor will have 3 conditions approximating each coefficient's general ability to differentiate between strong (the 0/.8 and 0/.3/.8 conditions), moderate (the 0/.5 and 0/.3/.5/.8) and a 'challenge' condition (.5/.8 and 0/.25/.33/.5/.66/.8).

The second condition will be the distribution of predictor variables. As noted in several instances of the statistical literature, two-group DDA has a 'common alternative' in logistic regression which uses group membership as the dependent variable and then models the log-likelihood of said membership. It has also been suggested that under conditions of multivariate normality and identical covariance matrices that DDA is preferable to logistic regression but that logistic regression is preferable when these assumptions are violated [Press and Wilson, 1978]. Because of this important distinction for a practical researcher, two conditions for the predictor variables will be varied: normal and non-normal. Again, the non-normal condition will build upon Finch's study and use the Fleishman method [1978] to simulate non-normal data with skew 1.75 and kurtosis of 3.75. It should also be noted that the distribution type was found to have significant interactions with correctly identifying variables associated with group separation in Finch's study.

The third condition will test the performance when the equality of the covariance matrices between groups is varied between equal and unequal (again mirroring Finch). Unequal covariance matrices will be created by simulating one group with 3x the value of the standard deviation as the other. Finch's study used a value 5x and this condition was found to have a marked effect. By dialing down the difference in the covariance matrices additional insight can be gained in this experiment by not only comparing sensitivity of each coefficient to the equality of covariance assumption but by also determining the degree to which a 3x and a 5x value between two groups might affect performance.

The fourth condition varied was number of predictors. The key studies of Huberty [1975], Barcikowski and Stevens [1975], Dalgleish [1994], Thompson [1991] and Finch [2008] showed some variability in this condition for their studies. Earlier studies by Barcikowski and Stevens were criticized for having too many to be practical by Thompson, while Huberty and Thompson generated *p*'s with 10 and 12 predictors respectively. Finch limited his study to 2 and 6 and concluded that these two conditions were 'very similar' though Type-1 Error rates were slightly lower in the six variable case. In that I suspect the two variable case may perform differently for the two coefficients when standardized coefficients are analyzed, I will keep the two variable condition and to maintain comparability between the studies I will also use Finch's condition of 6 predictors. Future work can analyze the performance on an increasing number of predictors if the findings of this study merit further investigation.

The fifth condition was sample size. With the literature recommending between 20 [Stevens, 2002] and 40 [Barcikowski and Stevens 1975] cases per predictor variable, this study will vary four conditions: 40, 80, 120 and 240. These differ somewhat from Finch's study but follow his general direction of including small to medium sized studies.

Finally, while Finch did vary the correlation among predictor variables, this study will not include this as a condition given the large number of factors already in play. An overall moderate correlation of .3 will be assumed for the 2 and 6 predictor variables – a number reflective of likely research scenarios and more realistic than assuming no correlation.

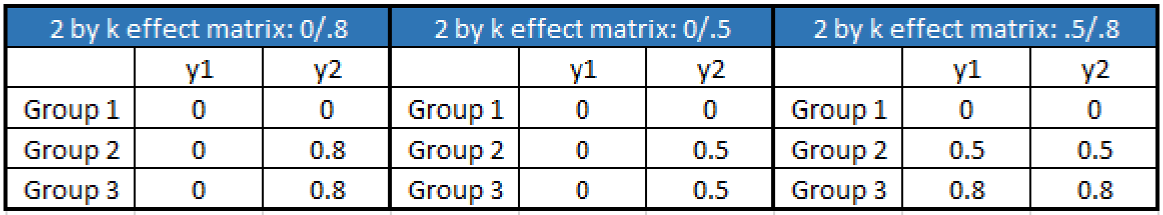
The resulting 3 x 2 x 2 x 2 x 4 (3 effect size scenarios by 2 distribution types by 2 equality of covariance assumptions by 2 predictors by 4 sample sizes) results in 192 conditions. 1,000 data sets will be simulated for each condition and 1,000 bootstrapped datasets (sampled with replacement) will then be created for each replication. The single linear discriminant function will be estimated for each bootstrapped dataset by the lda() function from Ripley's MASS package in R using maximum likelihood estimation (method = 'mle'). The structure-Rs and standardized coefficients will be captured and stored and a bootstrapped mean and standard deviation will be constructed for both. Given the slightly superior performance of the percentile bootstrap in Dalgleish's study over the standard bootstrap, the percentile bootstrap (2.5%, 97.5%) will be used to estimate Power for the effects simulated to be nonzero and Type-1 Error for effects simulated to be zero. Traditional cut scores of .3, .4 and .5 will also easily be evaluated though expectations are not high for this method due to its poor performance in Finch's study.

Conditions for the Study Comparing Contributions to Differences between Three Groups

The second study will operate under the three group condition and will be more restrained in its design. Finch's study in 2008 found broadly that 1) the percentile bootstrap was the preferred method for interpreting structure-Rs when sample sizes were greater than 100 and 2) relying solely on structure-Rs 'oversimplifies' the process. Finch's resampling techniques were rooted in Dalgleish's 1994 study which found the percentile bootstrap and standard bootstrap to be the most effective methods in interpreting group differences in a three group DDA. However, Dalgleish's data generation approach for the three group case was somewhat limited with regards to broader generalizations for methodologists. In his study, Dalgleish extracted the structure-R parameters and covariance matrix of an actual DDA study and invoked a SAS macro to generate 1,000 data sets based on the canonical correlations for each discriminant function and structure-Rs for the entire correlation matrix. When the structure-Rs did not 'align' with the generated population data Dalgleish discarded the data set which may have weakened the generalizability of the study. As stated before, Dalgleish also dismissed the standardized coefficients from his study. Finally, there was no discussion of whether Dalgleish's seed data was multivariate normal or had equal covariance matrices: critical assumptions that have been shown to influence DDA results. Interestingly, Thompson's Monte Carlo study laid the groundwork for an empirical comparison of the standardized coefficients and structure-Rs but compared only the mean deviance of the 1,000 generated data sets. So, a compact study blending the bootstrapping approach of Dalgleish [1994] and the data generation approach of Thompson [1991] and Huberty [1975] would inform the literature by extending the two-group comparison of the standardized coefficients and structure-Rs to the three group case.

This second study will generate 1000 multivariate normal data sets per condition for three groups using the mvrnorm() function in Ripley's MASS package for R. The function allows the specification of a vector of means according to a pre-specified covariance matrix. Matrix decomposition in the mvrnorm() function is performed via an eigenvector decomposition rather than the Choleski decomposition [Ripley, 1987]. For simplicity, a common covariance matrix assuming 'modest' correlations between the variables will be specified similar to Thompson in (1991) and Huberty (1975). The initial matrix will be populated with normal scores with mean of zero. Effect sizes will be achieved following Huberty's 1975 methodology that constructed a (p x k) population weight matrix from which a population mean matrix can be obtained. For each generated data set 1000 bootstrapped datasets will be created and scaling coefficients for each LDF (2 for this design) and the proportion of discriminatory power of each LDF as calculated by [10] will each be captured for each replication. From these datasets the standardized coefficients will be calculated for each bootstrapped data set via equation [14] and then the structure-Rs via $SC_w = R_wD$. In keeping with many methodological findings only the within correlations will be used to construct the structure-Rs. Trial runs of this experiment have suggested that the first LDF accounts for roughly 78% to 82% of the discriminatory power and if this holds true, only the first LDF will be analyzed in keeping with Huberty [1975]. The number of variables in the three group simulation will be limited to 2 and 6 and composes the first condition of the study. Other conditions shown to impact the performance of multi-group DDA will be varied as follows.

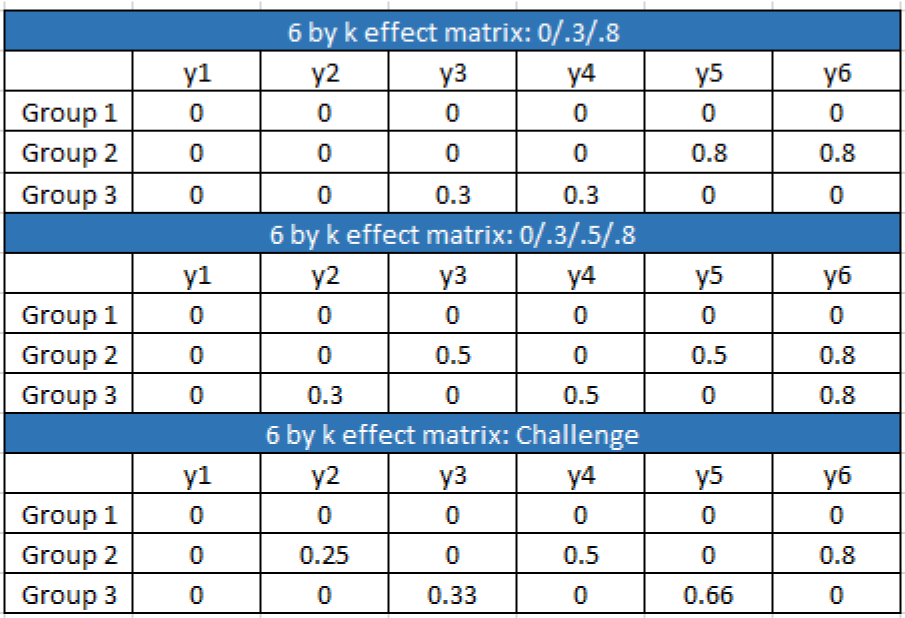
The second condition that will be varied is effect size. Similar to the proposed two-group design, three effect sizes will be constructed representing moderate, strong and 'challenging' group separation. Finch utilized Cohen's D in 2008 when varying effect sizes between .5 and .8, both values representing univariate measures that become more problematic when analyzing the three group case. Thompson's 1991 study utilized correlations of .1, .25 and .6 when constructing his population correlation matrices. Given that Finch's values are somewhat commensurate with Thompson's work I will construct the 'population weight' matrix using only .5 and .8 in the manner proposed in the tables below for both the two variable and six variable case.



2 by k effect matrix: 0/.8		
	y1	y2
Group 1	0	0
Group 2	0	0.8
Group 3	0	0.8

2 by k effect matrix: 0/.5		
	y1	y2
Group 1	0	0
Group 2	0	0.5
Group 3	0	0.5

2 by k effect matrix: .5/.8		
	y1	y2
Group 1	0	0
Group 2	0.5	0.5
Group 3	0.8	0.8



6 by k effect matrix: 0/.3/.8						
	y1	y2	y3	y4	y5	y6
Group 1	0	0	0	0	0	0
Group 2	0	0	0	0	0.8	0.8
Group 3	0	0	0.3	0.3	0	0

6 by k effect matrix: 0/.3/.5/.8						
	y1	y2	y3	y4	y5	y6
Group 1	0	0	0	0	0	0
Group 2	0	0	0.5	0	0.5	0.8
Group 3	0	0.3	0	0.5	0	0.8

6 by k effect matrix: Challenge						
	y1	y2	y3	y4	y5	y6
Group 1	0	0	0	0	0	0
Group 2	0	0.25	0	0.5	0	0.8
Group 3	0	0	0.33	0	0.66	0

These three general effect sizes are modeled after the two group case and will test each coefficient's ability to differentiate between strong (0/.8 and 0/.3/.8) effects, moderate effects (0/.5 and 0/.3/.5/.8) and a more difficult 'challenge' condition.

The third condition will be sample size. The same four sample sizes will be used in this experiment as in the two-group case as the recommendations for number of observations per variable do not increase in the literature as the number of groups is increased. Therefore, this condition will vary with 40, 80, 120 and 240 as the sample size values.

Finally, it should be noted that I am not proposing to test each coefficient's performance in lieu of violations of two important assumptions of three-group DDA: multivariate normality and the equality of covariance matrices. The reasons for this are two-fold. First, the complexity of simulating non-normal multivariate data with differing covariance matrices is a complex computing problem that is just now receiving attention in the literature [Mair, Satorra, Bentler, 2012]. Second, given the increased complexity of the three group case I thought it better to limit the number of conditions initially to determine if there was differential performance under this optimal case and leave the testing of the assumptions for a later study that could build on the foundations of this research.

The resulting 2 x 3 x 4 (two p values by three effect sizes (strong, medium and 'challenge') by 4 sample sizes (40, 80, 120 and 240) results in 24 conditions. 1000 data sets will be generated as described above using the mvrnorm() function and 1000 bootstrapped dataframes will then be constructed. The results of the study will be evaluated as described in the two group case: bootstrap estimates of the mean $\hat{\theta}_B$ and standard deviation $\hat{\sigma}_B$ will be compiled and the percentile bootstrap CI will be determined. Power and Type-1 Error will be assessed and sample results table will be presented below. Finally, as in the two-group case, cut scores of .3, .4 and .5 will also be evaluated.

Expected Results

Put succinctly, I expect Rencher's analytical conclusion that standardized coefficients better convey the multivariate nature of the data to be supported by these studies. Bootstrapping will allow the statistical testing of both the standardized coefficients and the structure-Rs by providing a mean estimate of each and its standard error for each condition [Dalgleish 1994]. This will allow a standard two-tailed Z-test under the null hypothesis that each variable does not contribute to group separation (or that the bootstrapped coefficient is equal to zero). Formally, this is expressed below for each coefficient:

$$\boldsymbol{H_0}\colon \hat{\theta}_{str.r_B} = 0$$

$$\boldsymbol{H_1}\colon \hat{\theta}_{str.r_B} \neq 0$$

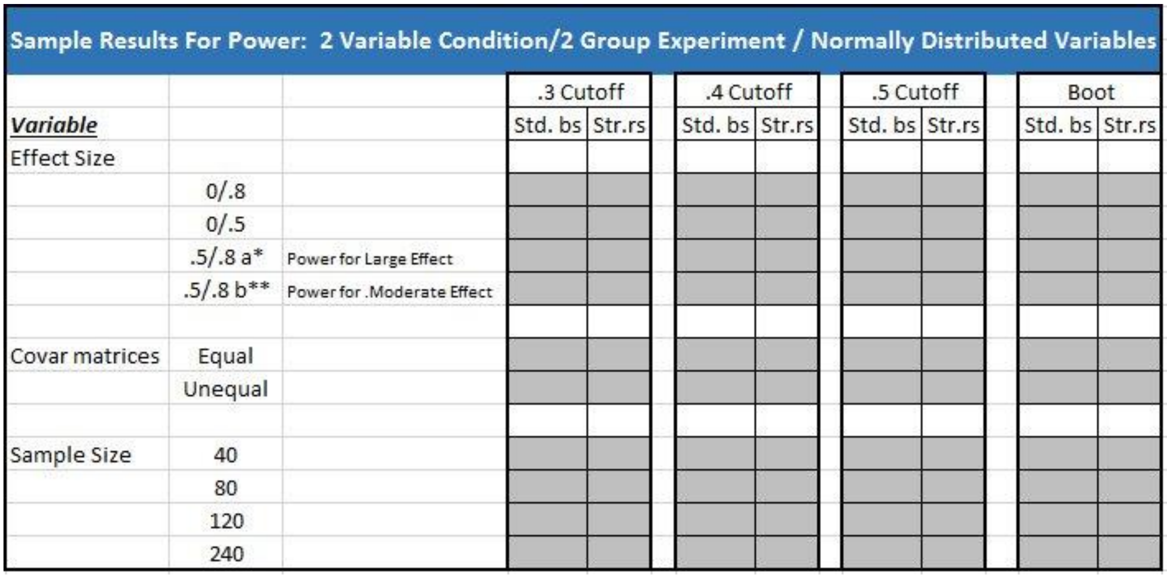
$$\boldsymbol{H_0}\colon \hat{\theta}_{std.b_B} = 0$$

$$\boldsymbol{H_1}\colon \hat{\theta}_{std.b_B} \neq 0$$

Several interesting questions can be pursued with the bootstrapped estimates. First, the efficacy of the cutoff values of .3, .4 and .5 can be judged by recording the number of times an effect designed to be greater than zero had bootstrapped coefficient values greater than the cutoff. Of course, power of the cut scores can be assessed as well. In this scenario, I expect that cut scores formed from the standardized coefficients will have better performance than those created from the structure-Rs. Second, statistical significance testing of each of the coefficients can be performed using the bootstrap percentile confidence interval and traditional two-tailed Z-testing. This will allow the explicit comparison of the sampling variability of the two coefficients at small and moderate sample sizes and should answer the question of whether one coefficient 'generalizes' better than the other. I expect to find that in fact their generalizability is either comparable or that standardized coefficients outperform (slightly) the structure-Rs.

The gist of the study will lie in comparing power and Type-1 Error for my effect size combinations – which have been explicitly set up to determine if standardized coefficients better convey the multivariate nature of the variables. Multiple relationships in the six variable case have been set to zero, and I expect the standardized coefficients have better Type-1 Error performance in these scenarios. Table C presented below provides a sample chart that will be populated with the percentage of bootstrapped estimates that incorrectly detected a significant effect for the zero coded variables. Power will also be calculated though I have fewer expectations with regards to these findings. Table A and B listed below provides a proposed table of results that will tabulate the percentage bootstrapped estimates that correctly identified the strong, moderate and small effects for the nonzero effect size variables in both the normal and non-normal condition for the two group case. Tables for the three-group case will be similarly constructed albeit with fewer conditions as explained in the three group research design. Finally, each 'challenge' condition was explicitly set up so that the variables can be ranked according to their cumulative increasing effect sizes. The percentage of simulations in which the rankings were correct will be directly compared for the two coefficients. It is here that I expect the standardized coefficients to significantly outperform the structure-Rs.

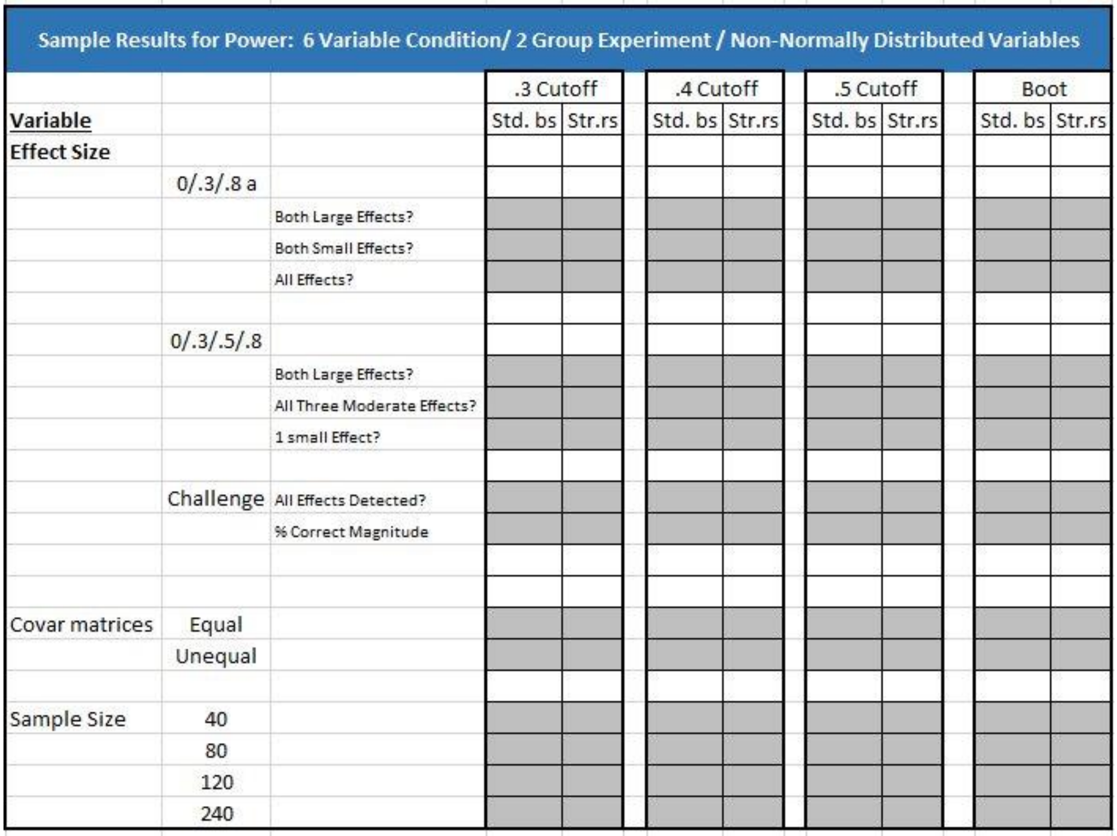
Sample Table A:



Sample Results For Power: 2 Variable Condition/2 Group Experiment / Normally Distributed Variables

			.3 Cutoff		.4 Cutoff		.5 Cutoff		Boot	
Variable			Std. bs	Str.rs	Std. bs	Str.rs	Std. bs	Str.rs	Std. bs	Str.rs
Effect Size										
	0/.8									
	0/.5									
	.5/.8 a*	Power for Large Effect								
	.5/.8 b**	Power for .Moderate Effect								
Covar matrices	Equal									
	Unequal									
Sample Size	40									
	80									
	120									
	240									

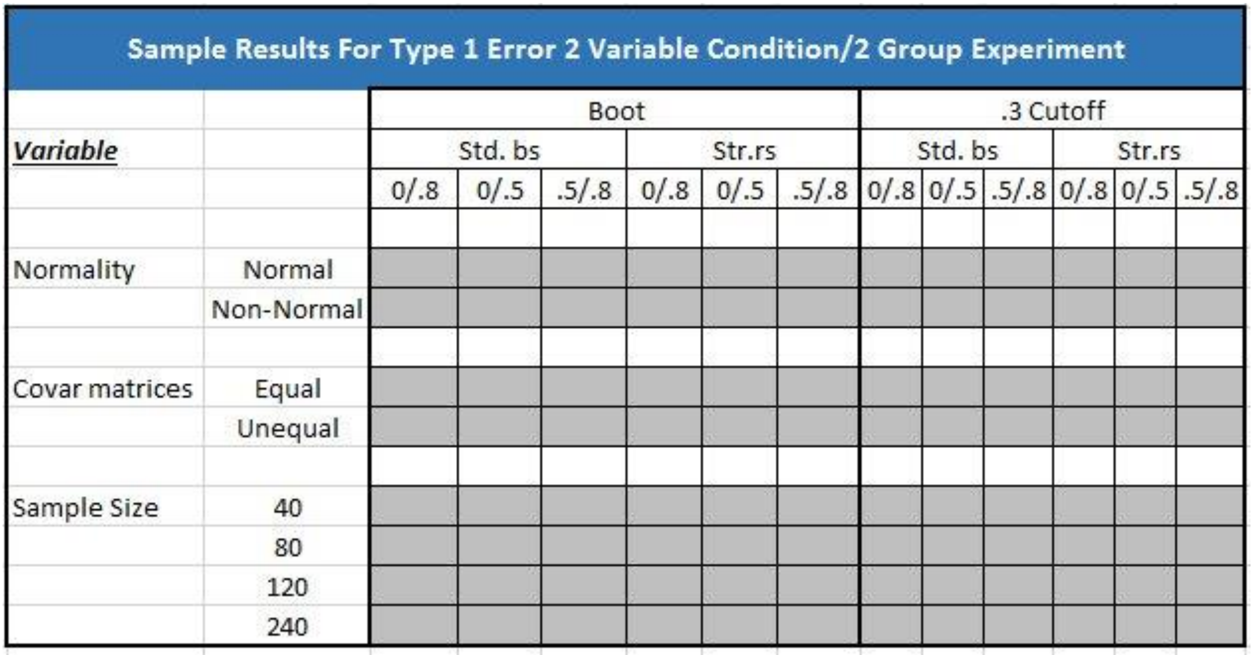
Sample Table B:



Sample Results for Power: 6 Variable Condition/ 2 Group Experiment / Non-Normally Distributed Variables

			.3 Cutoff		.4 Cutoff		.5 Cutoff		Boot	
Variable			Std. bs	Str.rs	Std. bs	Str.rs	Std. bs	Str.rs	Std. bs	Str.rs
Effect Size										
	0/.3/.8 a									
		Both Large Effects?								
		Both Small Effects?								
		All Effects?								
	0/.3/.5/.8									
		Both Large Effects?								
		All Three Moderate Effects?								
		1 small Effect?								
	Challenge	All Effects Detected?								
		% Correct Magnitude								
Covar matrices	Equal									
	Unequal									
Sample Size	40									
	80									
	120									
	240									

Sample Table C:



Sample Results For Type 1 Error 2 Variable Condition/2 Group Experiment

		Boot						.3 Cutoff					
Variable		Std. bs			Str.rs			Std. bs			Str.rs		
		0/.8	0/.5	.5/.8	0/.8	0/.5	.5/.8	0/.8	0/.5	.5/.8	0/.8	0/.5	.5/.8
Normality	Normal												
	Non-Normal												
Covar matrices	Equal												
	Unequal												
Sample Size	40												
	80												
	120												
	240												

References

Barcikowski, R. S., James (1975). "A Monte Carlo Study of the Stability of Canonical Correlations, Canonical Weights and Canonical Variate-Variable Correlations." Multivariate Behavioral Research **10**(3): 353-364.

Clarkson, D. B. (1979). "Estimating the Standard Errors of Rotated Factor Loadings." Psychometrika **44**(3): 297-314.

Cohen, J. (1988). Statistical Power Analysis. Hillsdale, NJ, Erlbaum.

Dalgleish, L. I. (1994). "Discriminant Analysis: Statistical Inference Using the Jackknife and Bootstrap Procedures." Psychological Bulletin **116**(3): 498-508.

Dalgleish, L. I. and D. Chant (1995). "A SAS Macro for Bootstrapping the Results of Discriminant Analyses." Educational and Psychological Measurement **55**(4): 613-624.

Efron, B. G., Gail (1983). "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation." The American Statistician **37**(1): 36-48.

Finch, H. (2009). "Identification of Variables Associated With Group Separation in Descriptive Discriminant Analysis: Comparison of Methods for Interpreting Structure Coefficients." The Journal of Experimental Education **78**(1): 26-52.

Finch, H. L., T (2008). "Evaluation of the Use of Standardized Weights for Interpreting Results From A Discriminant Analysis." Multiple Linear Regression Viewpoints **34**(1): 19-34.

Finch, W. H. (2006). "Misclassification Rates for Four Methods of Group Classification: Impact of Predictor Distribution, Covariance Inequality, Effect Size, Sample Size, and Group Size Ratio." Educational and Psychological Measurement **66**(2): 240-257.

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics.

Fleishman, A. I. (1978). "A Method for Simulating Non-Normal Distributions." Psychometrika **43**: 521-532.

Hastie, T. T., Robert; Friedman, Jerome (2009). The Elements of Statistical Learning. New York, NY.

Hess, B., et al. (2001). "The Efficacy of two Improvement-Over-Chance Effect Sizes for Two-Group Univariate Comparisons under Variance Heterogeneity and Nonnormality." Educational and Psychological Measurement **61**(6): 909-936.

Huberty, C. J. (1975). "The Stability of Three Indices of Relative Variable Contribution in Discriminant Analysis." The Journal of Experimental Education **44**(2): 59-64.

Huberty, C. J. O., S (2006). Applied MANOVA and Discriminant Analysis. Hoboken, NJ.

Huberty, C. J. W., Joseph (1992). "Variable Importance in Multivariate Group Comparisons." Journal of Educational Statistics **17**(1): 75-91.

Joy, O. M. T., John O. (1975). "On the Financial Applications of Discriminant Analysis." The Journal of Financial and Quantitative Analysis, **10**(5).

Mair, P. S., Albert; Bentler, Peter (2012). "Generating Nonnormal Multivariate Data Using Copulas: Applications to SEM." Multivariate Behavioral Research **47**(4): 547-565.

Rao, R. C. (1948). "The Utilization of Multiple Measurements in Problems of Biological Classification." Journal of the Royal Statistical Society. Series B (Methodological) **10**(2): 159-203.

Rencher, A. C. (1988). "On the Use of Correlations to Interpret Canonical Functions." Biometrika **75**(2): 363-365.

Rencher, A. C. (1992). "Interpretation of Canonical Discriminant Functions, Canonical Variates, and Principal
Components." The American Statistician **46**(3): 217-225.

Rencher, A. C. (2002). Methods of Multivariate Analysis. New York, NY.

Rencher, A. C. and D. T. Scott (1990). "Assessing the contribution of individual variables following rejection of a multivariate hypothesis." Communications in Statistics - Simulation and Computation **19**(2): 535-553.

Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge, England, Cambridge University Press.

Ripley, B. D. (2013) boot: Bootstrap Functions (originally by Angelo Canty for S).

Ripley, B. D. (2013). MASS: Support Functions and Datasets for Venables and Ripley's MASS.

Scott, E. (1978). "On the Financial Applications of Discriminant Analysis: Comment." The Journal of Financial and Quantitative Analysis **13**(1): 201-210.

Stevens, J. P. (2009). Applied Multivariate Statistics For The Social Sciences. New York, NY, Routledge.

Stoll, R. R. (1952). Linear Algebra and Matrix Theory. New York, NY, Dover.

Tatsuoka, M. M. (1972). Multivariate Analysis in Behavioral Research In F. Kerlinger (Ed.) Review of Research in Education. Itasca, IL, Peacock.

Thomas, D. R. (1997). "A Note on Huberty and Wisenbaker's "Views of Variable Importance"." Journal of Educational and Behavioral Statistics **22**(3): 309-322.

Thompson, B. (1991). "Invariance of Multivariate Results: A Monte Carlo Study of Canonical

Function and Structure Coefficients." The Journal of Experimental Education **59**(4): 367-382.

Venables, W. N. R., B.D. (2002). Modern Applied Statistics with S. USA.

Wickham, H. (2012). "plyr: Tools for splitting, applying and combing data."

Wickham, H. (2013). ggplot2: An implementation of the Grammar of Graphics.