Comparison of the Usefulness of Within-Group and Total-Group Structure Coefficients for Identifying Variable Importance in Descriptive Discriminant Analysis Following a Significant MANOVA: Examination of the Two-Group Case

Mercedes K. Schneider
Ball State University

This simulation study compared proportions of two types of structure coefficients in descriptive discriminant analysis, those based upon the error matrix, and those based upon the total matrix, for two groups from different populations with identical covariance matrices. The expected finding that the structure coefficients based upon the error matrix might be more appropriate than those based upon the total matrix was not supported.

Descriptive discriminant analysis (DDA) is a *post hoc* procedure useful for understanding the relationships among continuous variables following a significant MANOVA (Stevens, 2002; Tabachnick & Fidell, 2001). In DDA, linear discriminant functions (LDFs) are formed by weighting the p continuous variables such that separation of the k groups on the grouping variable is maximized. The number of LDFs possible is the smaller of p and $k - 1$. Thus, where there is one grouping variable with two levels, as is the case with the current study, only one LDF is possible.

Initially, a vector of raw weights (**v**) is calculated for a given LDF (Tatsuoka, 1988a, 1988b). However, these raw weights are not useful for interpretation. Instead, one commonly used coefficient for interpretation of LDF variable importance is the structure coefficient (SC), a measure of correlation between a given p and the associated LDF. The SC can be calculated using either the total group intercorrelation matrix (**R**) or the pooled within-group intercorrelation matrix (**W**) (Cooley & Lohnes, 1971; Huberty, 1975). These will be referred to as *total SCs* and *within SCs*, respectively. As part of the DDA output, SAS provides both the total and within SCs, and SPSS, only the within SCs. The p x 1 vector of total SCs ($\mathbf{s_T}$) for the only LDF in the two-group case may be calculated as follows:

$$\mathbf{s_T} = \mathbf{R}\mathbf{D}^{.5}\mathbf{v}\theta^{-.5} \quad (1)$$

where **D** is the diagonal matrix formed from the diagonal elements of $(1/N\text{-}1)\mathbf{T}$ (i.e., by multiplying the total SSCP matrix **T** by the reciprocal of the total sample size minus one); θ is the grand variance, and **R** and **v** are as defined previously (Cooley & Lohnes, 1971). The calculation of the p x 1 vector of within SCs ($\mathbf{s_W}$) follows the same formula, with the exception that the **W** matrix correspondingly replaces the **T** matrix.

Purpose of the Study

In his study of the ranking of LDF variable importance in DDA, Huberty (1975) notes that total SCs are appropriate if data "are considered representative of a single population" (p. 60) and within SCs, "if the underlying model is one of k populations with identical covariance matrices" (p. 60). As it appears no study has tested these stipulations, such was the primary goal of the current work. Specifically, the condition of $k = 2$ populations with identical covariance matrices was simulated, and the total and within SCs examined based upon criteria outlined in the procedures section to determine under what conditions within SCs might be better suited to total SCs in interpreting relative variable importance in DDA.

Procedures

This Monte Carlo simulation was executed using PROC IML in SAS. Two p-dimensional, multivariate population matrices were generated, with each being $N(\mu, \Sigma)$ (SAS Institute, 1999). The general procedure was as follows: In all cells, μ_1 was a p x 1 null vector, and μ_2, a p x 1 vector of effects of some combination such that $\mu_1 \neq \mu_2$. A sample of dimension n x p was then drawn from each population ($n_1 = n_2$; $p_1 = p_2$) and analyzed as a two-way MANOVA using Wilks' Λ and a special case of Bartlett's V as a test of significance:

$$V = -[N - 1 - (p + 2) / 2] \ln \Lambda \quad (2)$$

where Λ is also calculated using a modified formula

$$\Lambda = 1 / (1 + \lambda) \qquad (3)$$

since in the two-group analysis, λ is the only characteristic root. The special case of Bartlett's V is approximately a χ^2 distribution with *p* degrees of freedom (Tatsuoka, 1988a, 1988b).

Specifically, the variables and corresponding levels manipulated in this study were as follows:

1. $p = 2, 3$, and 4.
2. $n = 10, 50, 100$, and 500.
3. The population correlation matrices, $\mathbf{P}_1$ and $\mathbf{P}_2$. Five levels were used, reflecting five possible ranges of *p* intercorrelation: 0 - .20; .21 - .40; .41 - .60; .61 - .80, and .81 – 1.00. For any given experiment, the exact correlation for the two groups between continuous variables *p* and *p'* (where $p \neq p'$) was randomly generated within any one of these five ranges. The two most highly correlated ranges were included to investigate the effects of collinearity upon $\mathbf{s_T}$ and $\mathbf{s_W}$.
4. Population mean vector, μ_2. As previously mentioned, μ_1 was held constant as a null vector. Thus, μ_2 was manipulated as the vector of effects. The *p* elements of a given μ_2 were some combination of effects in standard deviations, with three possible levels of standard deviation used: 0, .5 and 1. These levels were arbitrarily selected to represent a three-tiered conceptualization of relative variable influence: *negligible*; *moderately influential*, and *highly influential* respectively. This manner of ranking variables will be discussed in the following section. In addition to its usefulness in defining a variable with a negligible contribution, the difference of 0 SD was included to investigate the influence of noncontributing variables upon $\mathbf{s_T}$ and $\mathbf{s_W}$. In sum, 20 *p* x 1 mean vector pairs were analyzed: 5 for $p = 2$; 7 for $p = 3$, and 8 for $p = 4$. (See Tables 1 thru 20 for the specific μ_2 investigated for a given cell.)

Each *n* x *p* cell was replicated 5,000 times. For the replications where the MANOVA null hypothesis $\mathbf{H_0}$**:** $\mu_1 = \mu_2$ was correctly rejected within each cell, the *p* x 1 vectors of total and within SCs, $\mathbf{s_T}$ and $\mathbf{s_W}$, were calculated. An expected pattern across $\mathbf{s_T}$ and $\mathbf{s_W}$ was identified for each mean effect vector $\underline{\mu_2}$ and the proportions of $\mathbf{s_T}$ and $\mathbf{s_W}$ vectors conforming to the identified pattern calculated. These proportions of $\mathbf{s_T}$ and $\mathbf{s_W}$, vectors were then compared across all levels of *n*, *p*, and **P**. The use of identified SC patterns will be discussed shortly.

Three-tiered Ranking of Relative Variable Importance

It may be tempting to apply some absolute criterion to SCs in order to make the determination of a corresponding continuous variable's contribution to group separation in DDA. Pedhazur (1997) cites the general guideline that an SC value of .3 is "meaningful" (p. 934). However, this guideline is too general to frame even a three-tiered ranking of relative variable importance upon separation of the groups. At the opposite end of the ranking spectrum are studies in which the SC values of a large number of variables are ranked in order of size without consideration of a more relaxed ranking system, such as the three-tiered system proposed above. For example, Huberty (1975) compared the utility of three types of weights/coefficients (total SCs, within SCs, and standardized weights) for variable ranking where the number of groups $k = 3, 4$, and 5, and the number of variables was held constant at $p = 10$. In essence, such rankings amounted to a ten-tiered ranking system. Generally, all coefficients/weights fared poorly regarding the correct ranking of the 10 variables. These results agreed with the findings of Barcikowski and Stevens (1975), who also investigated the utility of standardized weights and structure coefficients for ranking variable importance in canonical correlation. In this latter study, the simplest canonical dimension examined involved two variables in one variate and five, in the other variate. Thus, the smallest number of continuous variables investigated in the Barcikowski and Stevens work was $p = 7$, and in a more complex analysis than DDA, the canonical correlation (Tabachnick & Fidell, 2001). Both Huberty, and Barcikowski and Stevens note that a large sample size would be necessary for positive results for a 1-to-*p* ranking system (e.g., a 42:1 to 68:1 *n:p* ratio in Barcikowski & Stevens). Thus, with the minimum number of variables $p = 7$, the total sample size needed for utilizing SCs in DDA would be $N = 294$ (42:1) or $N = 476$ (68:1). As an additional objective of this study, the group sample size *n* will be investigated in conjunction with using a more relaxed ranking system.

Investigation of Levels of p *= 2, 3, and 4*

Rather than generally discounting SC usefulness for relative variable importance in DDA based upon studies employing large numbers of continuous variables (Barcikowski & Stevens, 1975; Huberty, 1975), it appears useful to investigate the more basic multivariate analyses, such as two-group DDA with $p = 2$, 3, and 4 variables, for two reasons. First, in educational research, a smaller number of p variables more realistically reflects that which is currently investigated in multivariate research (Schneider, 2002). Second, given the necessity of extremely large sample sizes in DDA where p is large (Stevens, 2002), limiting the number of continuous variables appears logical when resources, such as the number of participants in a study, are limited.

Identified SC Patterns

The identified pattern for a given mean vector effect μ_2 was chosen by examining the *range* of SC values corresponding to each element of μ_2 when $n = 1000$ across both types of SCs and all levels of **P**. A vector of least restrictive ranges for each of the p elements was then constructed to represent the identified pattern of ranges of SCs for the corresponding mean effect vector μ_2. The decision to use $n = 1000$ for identifying SC ranges is based upon the author's prior experience. When the group size is $n = 1000$ in two-group DDA, the SCs are sufficiently stabilized for interpretation of relative variable importance given the three-tiered ranking. Specifically, the ranges of SCs reflecting the effects of 0, .5, and 1 SD, respectively, are mutually exclusive. Increasing the group size n to 5000, for example, would have produced narrower ranges; however, such adjustment was not deemed necessary to term a coefficient value as indicating a ranking so general as *moderately influential*, for example.

A second reason for identifying SC vector patterns in the form of ranges has to do with the nature of the SCs in general. The description of the SC as a measure of correlation between a variable and the associated LDF might lead one to believe that the value of a given element is not influenced by the values of other elements in the vector. This is not the case. In DDA where the number of groups $k = 2$, the squared elements (e^2_i 's) of an SC vector, whether $\mathbf{s_T}$ or $\mathbf{s_W}$, must sum to one:

$$e^2_1 + e^2_2 + \ldots + e^2_p = 1\,, \tag{4}$$

where there are p elements. (Note: If $(k - 1) \geq p$, the proportions of *trace* of the correlation matrix, either **R** or **W**, as defined previously, accounted for by the p SC vectors would sum to one [Cooley and Lohnes, 1971]). This condition has implications for assigning unchanging SC values to indicate an effect of, say, 1 SD in the effect vector, μ_2. A variable on which the two groups differ by 1 SD will have a higher SC value in a vector of dimension p x 1 if the remaining $p - 1$ variables differ by .5 SD than if the remaining $p - 1$ variables also reflect a difference between groups of 1 SD. For example, compare the elements of the two 4 x 1 total SC ($\mathbf{s_T}$) vectors in Illustration 1 below. These two vectors have been written in the form of the condition outlined in Equation 4 above. In each, the last, bolded element is the referent kept constant at an effect of 1 SD. The remaining 3 elements represent an effect of .5 SD in the first vector and 1 SD in the second vector.

$$.3880698^2 + .4231913^2 + .4083566^2 + \mathbf{.7096167}^2 = 1.0000000 \tag{5}$$
$$.5021745^2 + .5076089^2 + .4923756^2 + \mathbf{.4977151}^2 = 1.0000001$$

Though both bolded coefficient values above represent a population effect of 1 SD between groups on the last variable in a set of 4, one can readily see the problem with applying some absolute, unchanging criteria value to indicate an effect of 1 SD. Instead, it is useful to note that the coefficient values in the first LDF above may indicate a present-but-lesser influence for the first three variables on group separation (coefficients in some expected range of .25 to .47, for example) and a prevalent influence for the last coefficient (which might be considered to be within some expected range of .68 to .82).

Table 1. Proportion of Total and Within SCs Fitting the Identified Pattern

Level of *n*	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.3188	.2825	.2881	.2475	.2289
	.3710	**.3441**	**.3438**	**.2913**	**.2473**
50	.5007	.4991	.4955	.4833	.4925
	.5209	**.5182**	**.5165**	**.5020**	**.5083**
100	.6248	.6284	.6377	.6251	.6326
	.6372	**.6450**	**.6505**	**.6418**	**.6484**
500	.9486	.9472	.9442	.9430	.9398
	.9530	**.9542**	**.9510**	**.9476**	**.9438**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 0 \\ .5 \end{bmatrix}$

Identified SC Range:

Absolute value of $\begin{bmatrix} 0-.25 \\ .96-1.0 \end{bmatrix}$

Table 2. Proportion of Total and Within SCs Fitting the Identified Pattern

Level of *n*	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.2565	.2397	.2356	.2127	.2111
	.3147	**.3035**	**.2909**	**.2560**	**.2447**
50	.4672	.4662	.4810	.4738	.4618
	.5178	**.5140**	**.5334**	**.5214**	**.5154**
100	.6146	.6250	.6232	.6306	.6238
	.6694	**.6756**	**.6814**	**.6830**	**.6818**
500	.9516	.9516	.9514	.9518	.9504
	.9702	**.9706**	**.9676**	**.9696**	**.9698**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

Identified SC Range:

Absolute value of $\begin{bmatrix} 0-.14 \\ .99-1.0 \end{bmatrix}$

Interpretation of the LDFs in Illustration 1 introduces another important point. Because of the condition imposed upon SCs as shown in Equation 4, it is important to note that the three-tiered ranking is only completely evident when all three of the SD effects are represented in the effect vector μ_2. For effect vectors with only two of the three SD levels present (e.g., .5 SD and 1 SD, as is the case in the first LDF in Illustration 1), only a two-level means of relative comparison is possible (i.e., *less influential/negligible*, or *more influential*, respectively). Finally, when all elements of a given effect vector $\underline{\mu_2}$ are the same (e.g., all are 1 SD, as is shown in the second LDF in Illustration 1), one "ranking" is possible (i.e., *all variables are contributing equally*). Thus, these two lower levels of variable ranking are subsumed in the three-tiered ranking system. Only if the SC values were unchanging could the three-tiered ranking system remain unaffected by the absence of one (or two) of the three SD levels from a given effect vector μ_2.

Results

Tables 1 through 20 present the results of the 20 mean effect vectors ($\underline{\mu}_{2i}$, where i = 1 - 20) examined in this study. For each μ_2, a vector of corresponding SC ranges was identified. This identified pattern is included in each table. Also included are the proportions of total- and within-SCs ($\mathbf{s_T}$ and $\mathbf{s_W}$, respectively) conforming to the identified pattern for each *n* x **P** cell. The $\mathbf{s_T}$ and $\mathbf{s_W}$ proportions were compared within each *n* x **P** cell, with the larger proportion interpreted as indicating greater usefulness of the referent SC (either $\mathbf{s_T}$ or $\mathbf{s_W}$).

Tables 1 through 5 include the results for the five mean effect vectors where the number of continuous variables $p = 2$. In general, $\mathbf{s_W}$ proportions were higher for vectors including elements with no effect (0 SD) (Tables 1 and 2) and vectors where both variables contributed differently (Table 3). In contrast, for 2 x 1 effect vectors where both variables contributed equally (Tables 4 and 5), $\mathbf{s_T}$ proportions were generally higher than those of $\mathbf{s_W}$ (the exception being the cell with n = 500 and **P** > .8, where both proportions were 1.0).

Table 3. Proportion of Total and Within SCs Fitting the Identified Pattern

Level of n	Level of P				
	0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.2789 **.2711**	.2917 **.2958**	.3058 **.3276**	.3222 **.3592**	.3371 **.4001**
50	.4564 **.4604**	.4857 **.4965**	.5275 **.5406**	.6004 **.6504**	.7036 **.8178**
100	.6286 **.6414**	.6516 **.6744**	.6766 **.7106**	.7564 **.8220**	.8474 **.9170**
500	.9060 **.9542**	.9362 **.9696**	.9620 **.9876**	.9722 **.9908**	.9924 **.9984**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ 1 \end{bmatrix}$

Identified SC Range: $\begin{bmatrix} .34 - .54 \\ .84 - .94 \end{bmatrix}$

Table 4. Proportion of Total and Within SCs Fitting the Identified Pattern

Level of n	Level of P				
	0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.3258 **.2502**	.3674 **.2808**	.3744 **.2747**	.5899 **.4396**	.6177 **.4667**
50	.5023 **.4678**	.5341 **.5028**	.6119 **.5709**	.7041 **.6701**	.9370 **.9216**
100	.6365 **.6066**	.6957 **.6608**	.7713 **.7483**	.8749 **.8535**	.9230 **.9125**
500	.9458 **.9324**	.9844 **.9772**	.9842 **.9778**	.9974 **.9966**	1.000 **1.000**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ .5 \end{bmatrix}$

Identified SC Range: $\begin{bmatrix} .58 - .81 \\ .58 - .81 \end{bmatrix}$

Table 5. Proportion of Total and Within SCs Fitting the Identified Pattern

Level of n	Level of P				
	0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.3699 **.2614**	.4042 **.2779**	.4845 **.3438**	.5154 **.3491**	.7010 **.5013**
50	.6328 **.5300**	.6584 **.5514**	.7321 **.6268**	.8666 **.7752**	.9533 **.8961**
100	.7856 **.6878**	.8736 **.7822**	.9238 **.8452**	.9748 **.9430**	.9976 **.9876**
500	.9946 **.9750**	.9998 **.9924**	.9992 **.9942**	1.000 **1.000**	1.000 **1.000**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Identified SC Range: $\begin{bmatrix} .63 - .78 \\ .63 - .78 \end{bmatrix}$

Tables 6 through 12 include results for the seven mean effect vectors where the number of continuous variables p=3. At this juncture, interpretation of the proportions becomes more complicated. For the effect vector with two of three variables not contributing (Table 6), $\mathbf{s_W}$ proportions were higher. When the condition was changed to one variable not contributing instead of two (Table 7), the $\mathbf{s_T}$ proportions were higher for cells of smaller size and lesser intercorrelation (i.e., $n \leq 100$ and $\mathbf{P} \leq .8$), and the $\mathbf{s_W}$ proportions higher for the largest cells (n = 500) and the highest intercorrelation (**P**>.8). Finally, when the condition was changed to all variables contributing, and all equally (.5 SD) (Table 8), $\mathbf{s_T}$ proportions were higher for all cells except the n = 500 x **P**>.8 cell, where both proportions equaled 1.0).

Table 6. Patterns for both Total and Within SCs where p = 3 and k = 2

Level of n	Level of P: 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0811	.0676	.0841	.0646	.0750
	.1146	**.1152**	**.1164**	**.0886**	**.0847**
50	.2249	.2146	.2230	.2660	.3202
	.2466	**.2307**	**.2413**	**.2844**	**.3346**
100	.3313	.3158	.3366	.3528	.4470
	.3502	**.3337**	**.3581**	**.3744**	**.4660**
500	.8306	.8356	.8372	.8392	.8376
	.8456	**.8504**	**.8546**	**.8528**	**.8520**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 0 \\ 0 \\ .5 \end{bmatrix}$

Identified SC Range:

Absolute value of * $\begin{bmatrix} .00-.26 \\ .00-.26 \\ .94-1.0 \end{bmatrix}$

* Vectors were either all positive or all negative

Table 7. Patterns for both Total and Within SCs where p = 3 and k = 2

Level of n	Level of P: 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.1322	.1616	.1723	.2011	.2676
	.1115	**.1356**	**.1581**	**.1992**	**.2804**
50	.2834	.3125	.3675	.4757	.5226
	.2719	**.3034**	**.3555**	**.4744**	**.5300**
100	.4375	.5017	.5695	.6442	.7146
	.4264	**.4889**	**.5609**	**.6400**	**.7228**
500	.9352	.9654	.9748	.9768	.9712
	.9220	**.9598**	**.9748**	**.9778**	**.9728**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 0 \\ .5 \\ .5 \end{bmatrix}$

Identified SC Range:

Absolute value of * $\begin{bmatrix} .00-.22 \\ .58-.81 \\ .58-.81 \end{bmatrix}$

* Vectors were either all positive or all negative

Table 8. Patterns for both Total and Within SCs where p = 3 and k = 2

Level of n	Level of P: 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0933	.1282	.1941	.2637	.5446
	.0640	**.0531**	**.0990**	**.1348**	**.3822**
50	.2085	.2821	.4045	.5480	.8761
	.1805	**.2386**	**.3562**	**.4907**	**.8559**
100	.3614	.4253	.4790	.7315	.9388
	.3235	**.3781**	**.4390**	**.6952**	**.9291**
500	.8722	.8974	.9626	.9884	1.000
	.8422	**.8662**	**.9492**	**.9852**	**1.000**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ .5 \\ .5 \end{bmatrix}$

Identified SC Range:

$\begin{bmatrix} .46-.67 \\ .46-.67 \\ .46-.67 \end{bmatrix}$

Table 9. Patterns for both Total and Within SCs where p = 3 and k = 2

Level of n	Level of P: 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.1221	.1587	.1575	.2216	.2988
	.1022	**.1377**	**.1569**	**.2382**	**.3713**
50	.3497	.4043	.4697	.5151	.6368
	.3283	**.3914**	**.4657**	**.5282**	**.7016**
100	.5642	.6226	.6932	.7684	.7642
	.5266	**.5950**	**.6916**	**.7798**	**.7920**
500	.9506	.9660	.9782	.9866	.9934
	.9274	**.9454**	**.9582**	**.9704**	**.9810**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ .5 \\ 1 \end{bmatrix}$

Identified SC Range:

$\begin{bmatrix} .28-.59 \\ .28-.59 \\ .74-.86 \end{bmatrix}$

Table 10. Patterns for both Total and Within SCs where p = 3 and k = 2

Level of n	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0812 **.0529**	.1117 **.0754**	.1312 **.0848**	.1811 **.1524**	.2871 **.2702**
50	.2590 **.2004**	.3177 **.2639**	.3948 **.3319**	.5196 **.4621**	.6762 **.6222**
100	.4812 **.3960**	.5308 **.4532**	.6118 **.5356**	.7812 **.7148**	.8962 **.8504**
500	.9150 **.8688**	.9506 **.9168**	.9840 **.9658**	.9966 **.9902**	.9996 **.9976**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ 1 \\ 1 \end{bmatrix}$

Identified SC Range:

Absolute value of * $\begin{bmatrix} .25-.46 \\ .61-.73 \\ .61-.73 \end{bmatrix}$

* Vectors were either all positive or all negative

Table 11. Patterns for both Total and Within SCs where p = 3 and k = 2

Level of n	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0728 **.0340**	.1012 **.0437**	.1338 **.0520**	.1946 **.0698**	.6803 **.4621**
50	.2472 **.1606**	.3072 **.2024**	.4096 **.2776**	.5485 **.3995**	.7849 **.6484**
100	.4262 **.2952**	.5068 **.3608**	.6284 **.4628**	.7968 **.6552**	.9808 **.9472**
500	.9458 **.8494**	.9662 **.8884**	.9890 **.9502**	.9996 **.9930**	1.000 **.9998**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Identified SC Range:

$\begin{bmatrix} .52-.63 \\ .52-.63 \\ .52-.63 \end{bmatrix}$

Table 12. Patterns for both Total and Within SCs where p = 3 and k = 2

Level of n	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0751 **.0861**	.0736 **.0861**	.0691 **.0896**	.0817 **.1019**	.0993 **.1184**
50	.2275 **.2467**	.2436 **.2597**	.2568 **.2862**	.3372 **.3942**	.3918 **.4618**
100	.3746 **.4014**	.3984 **.4338**	.4310 **.4828**	.4868 **.5556**	.5914 **.6676**
500	.8822 **.9158**	.9010 **.9420**	.9204 **.9586**	.9460 **.9702**	.9462 **.9716**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 0 \\ .5 \\ 1 \end{bmatrix}$

Identified SC Range:

Absolute value of $\begin{bmatrix} .00-.13 \\ .35-.54 \\ .84-.94 \end{bmatrix}$

In short, the dominant coefficient apparently shifted from $\mathbf{s_W}$ to $\mathbf{s_T}$ as the number of contributing variables was increased to the point that a vector of equally contributing variables was reached.

When the condition of all variables contributing equally at .5 SD (Table 8) was altered such that one variable contributed 1 SD toward group separation (Table 9), the dominant coefficient again shifted, this time with $\mathbf{s_T}$ proportions higher for **P** < .61 excepting cells where n = 500. When a second element in the effect vector was changed from .5 SD to 1 SD (Table 10), $\mathbf{s_T}$ proportions were consistently higher than $\mathbf{s_W}$; such remained true as the last of the three effects was increased from .5 SD to 1 SD (Table 11). Note that the condition in Table 11 was similar condition to that in Table 8, this time with all effect variables again contributed equally but with the greater effect of 1 SD. Thus, for the condition of all variables contributing, and contributing equally, $\mathbf{s_T}$ proportions were higher than $\mathbf{s_W}$.

One additional condition was examined where p = 3, that of all three variables contributing at the three different levels of effects (0, .5, and 1 SD) (Table 12). For this condition, $\mathbf{s_W}$ proportions were consistently higher than $\mathbf{s_T}$. It is interesting to note that $\mathbf{s_W}$ proportions are consistently higher for all three conditions where at least half of the *p* variables did not contribute to group separation (i.e., the effect equals 0 SD; see Tables 1, 2, and 11).

Tables 13 through 20 include the results of the eight conditions in this study involving as mean effect vectors $\boldsymbol{\mu}_2$ with $p = 4$ continuous variables. Unlike the somewhat systematic shifting of the predominance of one coefficient over another where the number of continuous variables $p = 3$, the vectors involving $p = 4$ are not so easily interpreted, generally speaking. Whereas for the two conditions with all mean effect vector elements contributing and contributing equally (Tables 14 and 18), the $\mathbf{s_T}$ proportions were consistently higher, for no condition examined were the $\mathbf{s_W}$ proportions consistently higher than $\mathbf{s_T}$. It should be noted that the number of conditions was limited and numerous possibilities omitted, including conditions where 2 or 3 elements in the 4 x 1 effect vector represented *p* variables contributing nothing toward group separation (that is, 2 or 3 elements in the effect vector $\boldsymbol{\mu}_2$ at 0 SD). Thus, the idea that $\mathbf{s_W}$ proportions might be consistently higher than $\mathbf{s_T}$ proportions where the number of noncontributing *p* variables equaled or exceeded half was not examined where $p = 4$.

The $\mathbf{s_W}$ proportions were higher than the $\mathbf{s_T}$ proportions for all but 4 of the 20 *n* x **P** cells for the condition where three effects were all set at .5 SD and one, at 1 SD (Table 15). These four cells having higher or equal proportions of $\mathbf{s_T}$ were four extreme cells (i.e., $n = 10$ and 500; $\mathbf{P} \leq .4$ and $\mathbf{P} > .6$). The $\mathbf{s_W}$ proportions were also higher for most of the cells where the mean effect vector $\boldsymbol{\mu}_2$ included one noncontributing variable (0 SD), two moderately contributing variables (.5 SD), and one predominantly contributing variable (1 SD) (Table 19). In Table 19, cells having higher $\mathbf{s_T}$ proportions tended to be those where the group size was very large ($n = 500$).

For three conditions, neither $\mathbf{s_W}$ nor $\mathbf{s_T}$ proportions were consistently higher, but $\mathbf{s_T}$ proportions were prevalent in the 20 *n* x **P** cells. The first such condition involved three of the four variables contributing equally at .5 SD and one, not contributing (0 SD) (Table 13). Here, $\mathbf{s_T}$ proportions were higher where $\mathbf{P} \leq .40$. As the level of intercorrelation among *p* increased, $\mathbf{s_W}$ was higher, initially for the largest group size $n = 500$ (**P** range: .41 to .60), then for the two largest group sizes $n = 100$ and 500 as the intercorrelation increased to the second highest range (**P** range: .61 to .80). Finally, where intercorrelation was at the highest level ($\mathbf{P} \geq .81$), $\mathbf{s_W}$ proportions were higher for the three largest levels of *n* ($n = 50$, 100 and 500). These findings are similar to those in the $p = 3$ condition with two variables contributing equally at .5 SD and one, not contributing (0 SD) (Table 7).

In the remaining two conditions, where two variables are contributing equally at .5 SD and two, contributing equally at 1 SD (Table 16) and one variable not contributing (0 SD), one contributing somewhat (.5 SD), and the remaining two, contributing equally at 1 SD (Table 20), $\mathbf{s_T}$ proportions were higher generally, with $\mathbf{s_W}$ proportions tending to be higher for greatest continuous variable intercorrelation ($\mathbf{P} \geq .81$) but not for any given group size *n*.

Discussion

Though the results for this study are sometimes complex to interpret, what is clear is that $\mathbf{s_W}$ proportions were not consistently higher than $\mathbf{s_T}$ proportions in this study where the two groups were generated from two populations with identical covariance matrices. From the cells investigated, one might expect $\mathbf{s_W}$ proportions to be higher when half of the *p* continuous variables are not contributing to group separation. It seems that this situation might likely occur when a researcher does what Stevens (2002) advises against in the context of MANOVA: including variables in an analysis without theoretical justification, or simply because the data were available. Too, when all variables are contributing equally, $\mathbf{s_T}$ proportions are consistently higher than $\mathbf{s_W}$. Finally, generally speaking, for conditions with mixed results (i.e., some cells with $\mathbf{s_T}$ proportions higher and some cells with higher $\mathbf{s_W}$ proportions), $\mathbf{s_W}$ proportions tended to be higher for greater *p*-variable intercorrelation ($\mathbf{P} \geq .81$) or larger group sizes *n* (i.e., $n = 500$).

Table 13. Patterns for both Total and Within SCs where $p = 4$ and $k = 2$

Level of *n*	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0621	.0723	.0872	.1120	.1576
	.0378	**.0513**	**.0605**	**.0913**	**.1507**
50	.1847	.2719	.2918	.3844	.5253
	.1682	**.2477**	**.2752**	**.3712**	**.5394**
100	.3704	.4144	.5179	.5998	.6992
	.3451	**.3911**	**.5033**	**.5998**	**.7150**
500	.9522	.9646	.9634	.9682	.9694
	.9450	**.9612**	**.9646**	**.9714**	**.9710**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 0 \\ .5 \\ .5 \\ .5 \end{bmatrix}$

*Identified SC Range: Absolute value of** $\begin{bmatrix} .00-.17 \\ .42-.70 \\ .42-.70 \\ .42-.70 \end{bmatrix}$

* Vectors were either all positive or all negative

Table 14. Patterns for both Total and Within SCs where $p = 4$ and $k = 2$

Level of *n*	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0254	.0524	.0814	.1424	.3699
	.0113	**.0143**	**.0287**	**.0528**	**.1781**
50	.0809	.1234	.1943	.3386	.6008
	.0663	**.0972**	**.1552**	**.2744**	**.5293**
100	.1956	.2590	.3688	.5595	.8771
	.1712	**.2199**	**.3266**	**.5099**	**.8572**
500	.7810	.8492	.9334	.9832	.9990
	.7336	**.8110**	**.9104**	**.9766**	**.9980**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ .5 \\ .5 \\ .5 \end{bmatrix}$

Identified SC Range: $\begin{bmatrix} .41-.59 \\ .41-.59 \\ .41-.59 \\ .41-.59 \end{bmatrix}$

Table 15. Patterns for both Total and Within SCs where $p = 4$ and $k = 2$

Level of *n*	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0439	.0555	.0577	.0902	.1238
	.0368	**.0510**	**.0630**	**.1287**	**.2098**
50	.1652	.2200	.2648	.3991	.5766
	.1654	**.2274**	**.2831**	**.4383**	**.6912**
100	.3610	.4220	.5060	.6212	.7708
	.3772	**.4534**	**.5392**	**.6808**	**.8226**
500	.8972	.9396	.9670	.9902	.9992
	.9364	**.9684**	**.9800**	**.9902**	**.9974**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ .5 \\ .5 \\ 1 \end{bmatrix}$

*Identified SC Range: Absolute value of** $\begin{bmatrix} .25-.47 \\ .25-.47 \\ .25-.47 \\ .67-.82 \end{bmatrix}$

* Vectors were either all positive or all negative

Table 16. Patterns for both Total and Within SCs where $p = 4$ and $k = 2$

Level of *n*	Level of P 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0244	.0413	.0500	.0621	.1036
	.0159	**.0266**	**.0275**	**.0636**	**.1173**
50	.1346	.1656	.2147	.3074	.5988
	.1074	**.1391**	**.1920**	**.3004**	**.6334**
100	.3076	.3684	.4840	.5784	.7664
	.2574	**.3348**	**.4422**	**.5554**	**.7534**
500	.9022	.9378	.9794	.9932	.9998
	.8694	**.9136**	**.9552**	**.9824**	**.9960**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ .5 \\ 1 \\ 1 \end{bmatrix}$

Identified SC Range: $\begin{bmatrix} .25-.41 \\ .25-.41 \\ .57-.71 \\ .57-.71 \end{bmatrix}$

Table 17. Patterns for both Total and Within SCs where *p* = 4 and *k* = 2

Level of *n*	Level of P: 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0353	.0534	.0734	.1335	.2462
	.0184	**.0269**	**.0408**	**.0735**	**.2267**
50	.1764	.2290	.3037	.3982	.6531
	.1234	**.1626**	**.2382**	**.3464**	**.6979**
100	.3690	.4492	.5732	.7252	.8248
	.2816	**.3566**	**.5038**	**.6952**	**.8658**
500	.9254	.9464	.9714	.9856	.9960
	.9056	**.9486**	**.9846**	**.9950**	**.9996**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} .5 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

Identified SC Range: $\begin{bmatrix} .20-.35 \\ .49-.62 \\ .49-.62 \\ .49-.62 \end{bmatrix}$

Table 18. Patterns for both Total and Within SCs where *p* = 4 and *k* = 2

Level of *n*	Level of P: 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0371	.0582	.1013	.1984	.4455
	.0131	**.0175**	**.0277**	**.0496**	**.1497**
50	.2132	.2842	.3558	.5222	.8764
	.1138	**.1600**	**.2126**	**.3397**	**.7762**
100	.4230	.5186	.6484	.8380	.9772
	.2628	**.3392**	**.4566**	**.6850**	**.9314**
500	.9732	.9834	.9982	.9998	1.000
	.8990	**.9304**	**.9842**	**.9986**	**1.000**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$

Identified SC Range: $\begin{bmatrix} .44-.56 \\ .44-.56 \\ .44-.56 \\ .44-.56 \end{bmatrix}$

Table 19. Patterns for both Total and Within SCs where *p* = 4 and *k* = 2

Level of *n*	Level of P: 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0191	.0273	.0335	.0436	.0819
	.0235	**.0267**	**.0387**	**.0459**	**.1170**
50	.1095	.1431	.1660	.2134	.3582
	.1055	**.1583**	**.1770**	**.2236**	**.4078**
100	.2456	.2886	.3664	.4632	.6256
	.2430	**.2936**	**.3870**	**.5054**	**.6684**
500	.8494	.8828	.9244	.9634	.9812
	.8180	**.8574**	**.8992**	**.9496**	**.9794**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 0 \\ .5 \\ .5 \\ 1 \end{bmatrix}$

Identified SC Range: Absolute value of * $\begin{bmatrix} .00-.15 \\ .34-.50 \\ .34-.50 \\ .74-.86 \end{bmatrix}$

* Vectors were either all positive or all negative

Table 20. Patterns for both Total and Within SCs where *p* = 4 and *k* = 2

Level of *n*	Level of P: 0 - .20	.21-.40	.41-.60	.61-.80	.81-1.0
10	.0224	.0213	.0348	.0428	.0776
	.0178	**.0213**	**.0262**	**.0343**	**.0746**
50	.1486	.1642	.2118	.2726	.4964
	.1364	**.1434**	**.1920**	**.2712**	**.5284**
100	.2934	.3456	.4458	.4816	.6938
	.2528	**.2994**	**.4126**	**.4558**	**.7112**
500	.9152	.9466	.9586	.9736	.9800
	.8596	**.9020**	**.9272**	**.9690**	**.9828**

Note: (SST SCs are in Roman print; SSW SCs, in **bold** print.)

Population Mean Vector: $\mu_2 = \begin{bmatrix} 0 \\ .5 \\ 1 \\ 1 \end{bmatrix}$

Identified SC Range: Absolute value of $\begin{bmatrix} .00-.11 \\ .27-.43 \\ .60-.72 \\ .60-.72 \end{bmatrix}$

* Vectors were either all positive or all negative

Issues of Practical Significance

Even though no significance tests were conducted regarding which of the two SC proportions were prevalent, the practical usefulness of the information available in the tables is worth comment. First, for a number of the conditions, the difference between proportions was minimal from a practical standpoint (i.e., ≈ .03 or less in many cells); thus, either $\mathbf{s_T}$ or $\mathbf{s_W}$ coefficients might be used for interpretation (see Tables 6, 7, 13, 16, 19, and 20). The primary goal of this study was to compare $\mathbf{s_T}$ and $\mathbf{s_W}$ coefficients to determine if $\mathbf{s_W}$ coefficients might be more useful where the two groups were from two populations with identical covariance matrices. Nevertheless, it may be useful to speak of three additional issues of practical significance though such issues are not directly related to the research goal. First, the SC values for variables not contributing to group separation (0 SD in the effect vector) tended to exhibit a range of approximately .00 to .26, and the signs of the elements in the resulting SC vectors could not be predicted (see Tables 1, 2, 6, 7, 12, 13, 19, and 20). As for the values of the coefficients of the noncontributing variables, such generally fits the rule cited in Pedhazur (1997) that coefficient values of .3 of above are useful.

Second, regarding use of the identified ranges to guide researcher interpretation, one can see that other factors must be considered when determining the group size necessary to achieve a given proportion in DDA, such as the effect between groups on each continuous variable and the degree of continuous variable intercorrelation. Moreover, given the jump in proportions between n = 100 and 500, in order for the tables of proportions to be useful, proportions associated with additional sample sizes between n = 100 and 500 must be examined. The inclusion of more mean effect values (e.g., .3 SD; .7 SD) might also further clarify the dynamic between higher $\mathbf{s_T}$ and $\mathbf{s_W}$ for conditions where neither coefficient consistently had a higher proportion.

Finally, whereas unusual, the use of identified SC patterns does offer the beginnings of a practical means of interpreting relative variable importance in DDA. As one compares the identified patterns with their respective effect vectors, one can see a relationship between the population effect vectors and sample SCs. How a researcher might utilize these identified patterns provides a focus for future DDA studies.

References

Barcikowski, R., & Stevens, J. P. (1975). A Monte Carlo study of the stability of canonical correlations, canonical weights and canonical variate-variable correlations. *Multivariate Behavioral Research, 10,* 353-364.

Cooley. W. W. & Lohnes, P. R. (1971). *Multivariate data analysis*. New York: John Wiley and Sons.

Huberty, C. J. (1975). The stability of three indices of relative variable contribution in discriminant analysis. *Journal of Experimental Education, 2,* 59-64.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace College Publishers.

SAS Institute (1999). *SAS/IML user's guide: Version 8.* Cary, NC: Author.

Schneider, M. K. (2002). *A Monte Carlo investigation of the Type I error and power associated with descriptive discriminant analysis as a MANOVA post hoc procedure.* Published doctoral dissertation, University of Northern Colorado.

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum.

Tatsuoka, M. M. (1988a). Multivariate analysis of variance. In J. R. Nesselroade and R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed.). New York: Plenium Press.

Tatsuoka, M. M. (1988b). *Multivariate analysis: Techniques for educational and psychological research* (2nd ed.). New York: Macmillan Publishing.

Tabachnick, B. G, & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Boston: Allyn and Bacon.

Send correspondence to: Mercedes K. Schneider, Ph.D.
Department of Educational Psychology
Ball State University, Teachers College 520
Muncie, IN 47306
Email: MSCHNEIDER@bsu.edu