Performance of the Roy-Bargmann Stepdown Procedure as a Follow Up to a Significant MANOVA

W. Holmes Finch
Ball State University

The Roy-Bargmann procedure has been suggested as a post hoc procedure for a significant MANOVA result. This method, which is based on application of the univariate General Linear Model, requires the researcher to order the dependent variables a priori in terms of their contextual importance. Subsequently, if the MANOVA is found to be statistically significant for a categorical independent variable, these response variables are tested individually using univariate analyses in the sequence established a priori. Thus, the first variable is treated as the dependent variable in an ANOVA and if the groups' means are found to differ significantly on this variable, it serves as a covariate while the second variable in the sequence is the response and the means of the groups are once again compared. This testing continues for each of the response variables, with variables higher in the sequence serving as covariates for those lower in importance. The current study was designed to examine the Type I error rate and power of this post hoc approach under a variety of data conditions. Results show that factors such as distribution of the dependent variables, equality (or lack thereof) of the covariance matrices and sample size all have a significant impact on Type I error and power. Furthermore, both Type I error rates and power of variables later in the sequence are influenced by variables earlier in the sequence.

Multivariate Analysis of Variance (MANOVA) is a popular tool used by social science researchers and others, allowing for the analysis of multiple dependent variables with one or more independent factors. The null hypothesis being tested by MANOVA is $\mu_1 = \mu_2 = \mu_3$ where μ_k is the vector of means for group *k*. When this hypothesis is rejected due to a significant test statistic, researchers may be interested in to which groups or dependent variable(s) the result applies. Given that rejection of the very general null hypothesis of the MANOVA indicates that there is some difference among the *k* groups on one or more of the *p* dependent variables. In order to gain a more complete understanding of the nature of such a significant effect, a researcher may want to use a follow up analysis designed to illuminate the significant result in terms of group differences on the response variables (Tabachnick & Fidell, 2007; Stevens, 1996). A number of such approaches have been discussed in the literature, including the Simultaneous Test Procedure (STP) (Gabriel, 1968), Descriptive Discriminant Analysis (DDA) (Huberty, 1994), a Step Down procedure (SD) (Roy, 1958), two groups multivariate comparisons (Stevens, 1972) and the use of univariate Analysis of Variance (ANOVA). It should be noted that with the exception of the latter approach, all of these methods retain the general multivariate flavor of the original analysis, albeit in very different ways. Indeed, several authors (e.g. Stevens, 1996) argue that whatever follow up to MANOVA is finally used, it needs to be based upon a multivariate platform. Nonetheless, most researchers who make use of MANOVA will have specific questions regarding the nature, in terms of both the response variables and the groups, of the significant differences signaled by the multivariate analysis. This study was designed to examine the performance of one of these follow up methods that may be effective in characterizing a significant MANOVA result.

Analysis of Variance (ANOVA)

Perhaps the most straightforward approach to investigating a significant MANOVA result is through the application of individual univariate ANOVA analyses for each of the dependent variables separately. This approach is facilitated by common statistical software packages such as SAS and SPSS, which print the univariate results with the multivariate. Despite this ease of use, the use of univariate ANOVA in this way has generally been rejected as a viable alternative for following up a significant MANOVA result because, as Enders (2003) points out, the univariate ANOVA does not accurately maintain the nominal Type I error rate in most cases (generally being too conservative), even when a correction such as Bonferroni or Holm is used. Indeed, Maxwell (1992) found that using such alpha corrections with univariate ANOVA to maintain the nominal experiment-wise Type I error rate only works when either the MANOVA null hypothesis is totally false, the MANOVA null hypothesis is totally true or the MANOVA null hypothesis is false for all but one of the dependent variables. In all other cases, using ANOVA to investigate a significant MANOVA will yield an incorrect Type I error rate. Keselman, Huberty, Lix, et al

(1998) assert that if the researcher is interested in a multivariate hypothesis then the follow up to a significant MANOVA should be multivariate in nature.

Two groups multivariate comparison

Another MANOVA follow up approach that has been suggested in the literature is the two groups multivariate comparison. As outlined by Stevens (1972), the two groups method involves using the Hotelling's T^2 test statistic (Hotelling, 1931) to compare all possible pairs of groups on the entire set of dependent variables simultaneously. As an example, if the MANOVA involved one independent categorical variable with 3 groups measured on 3 dependent variables, the two groups analysis would involve the multivariate comparison of groups 1 and 2, groups 1 and 3, and groups 2 and 3 on the 3 response variables simultaneously. Stevens (1972) compared this approach with DDA, SD and multivariate contrasts and found that the results, in terms of identifying group differences, were fairly similar, though the two groups approach did not allow for direct identification of group differences on individual variables. Thus, when two groups are found to differ the difference is for the entire set of responses, so that if a researcher would like to identify clearly for which of the response variables groups differ, (s)he could not do so using this method.

Simultaneous Test Procedures

Gabriel (1968) expanded upon earlier work of Roy and Bose (1953) to develop multivariate simultaneous confidence intervals for group differences using any one of the common MANOVA test statistics, such as Roy's greatest root, Pillai's Trace, Hotelling's trace and Wilks' Lambda. Various recommendations have been made regarding which of these is the most generally appropriate for use in the STP context (Sheehan, 1995; Elliott, 1993; Mudholkar, Davidson & Subbaiah, 1974; Olson, 1974). Simulation research on the power and Type I error control of the STP based on one or more of these statistics found that violations of the assumptions of normality and homogeneity of covariance matrices had a significant impact on their performance, such that no one approach could be identified as optimal in all cases (Sheehan, 1995; Bird & Hadzi-Pavlovic, 1983; Elliott, 1993). These studies also found that the more restrictive the hypotheses being tested, the lower the power of the STP procedure, regardless of the statistic used to form the confidence intervals. Indeed, Sheehan (1995) concluded that this approach was too conservative to investigate group differences on individual variables.

Descriptive Discriminant Analysis

Several authors have recommended the use of DDA as an appropriate follow up procedure for a significant MANOVA (Enders, 2003; Huberty, 1994; Tabachnick & Fidell, 2007; Stevens, 1996). The fact that DDA is a multivariate technique very closely related to MANOVA addresses concerns voiced by Keselman, et al. (1998) and Stevens (1996) that when the research questions of interest are multivariate in nature, then the analyses used to address these questions be multivariate as well. DDA involves the identification of a linear combination of the dependent variables that maximizes the differences among the groups, as expressed in the between and within sums of squares and cross products matrices, $\mathbf{S_B}$ and $\mathbf{S_W}$, respectively. These linear combinations of the original set of dependent variables can then be used to characterize the nature of the difference(s) among groups identified by the significant MANOVA. Interpretation of these discriminant functions can be carried out in at least two ways, using either Structure Coefficients (SC) (Huberty, 1994), which are values representing the correlations between individual observed variables and the overall discriminant function or standardized discriminant function coefficients (Rencher, 1992). While there is some disagreement as to which approach is preferable, in both cases the magnitudes of either of these values reflect the relative importance of each observed variable in defining the discriminant function, making them useful for identifying potentially important contributors (from among the set of dependent variables) to the observed differences among the groups, as well as providing some insight into the conceptual nature of the discriminant functions themselves (Enders, 2003; Huberty, 1994; Stevens, 1972; Rencher, 1992).

The fact that SC's may be used in practice to ascertain which of the observed variables are related to the one or more discriminant functions (Tabachnick & Fidell, 2007) makes the issue of their interpretation very important. By examining the magnitudes of SC's, a researcher can determine which of the dependent variables are most associated with the linear combination(s) that best differentiate the groups in question. Those with the largest SC's are deemed to be most associated with the group differences, and are thus interpreted more fully in terms of how they might differ from group to group. This difference is

often characterized by examining the means of the groups on those variables with sufficiently large SC values (Tabachnick & Fidell, 2007). A difficulty with employing this method is that there is not a formal hypothesis test available for the SC's, and thus various rules of thumb have been recommended as cutoffs for identifying "sufficiently large" values. Schneider (2004) conducted a simulation study using cut points of 0.3, 0.4 and 0.5 and found that the general ability of DDA to identify statistically meaningful variables, in terms of group differentiation, was lower for higher cut offs, except where the differences in group means was characterized by a large effect size. Schneider also found that the use of these various cutoff values led to the false identification of "important" variables at rates reaching as high as 0.8 in many cases. In other words, a researcher using these rules of thumb, would be likely to incorrectly conclude that particular variables were "important" in defining the discriminant function when they were in fact not.

Stepdown Analysis

The Stepdown analysis (SD) examined here was introduced by Roy (1958) and Roy and Bargmann (1958) and extended by others (Marden & Perlman, 1990; Mudholkar & Subbaiah, 1988; Kabe, 1984) and is based on the General Linear Model (GLM) in the form of Analysis of Covariance (ANCOVA). It has been recommended as a follow up procedure for a significant MANOVA by several authors, including Tabachnick and Fidell (2007), Stevens (1996), and Mudholkar and Subbaiah (1980). In addition, the SD technique can be used to derive a test of overall significance in the multivariate context as demonstrated by Kabe (1984), among others. The SD procedure involves a multi-step application of univariate linear models involving the dependent variables, much in the way that ANOVA does. However, there are some major differences between SD and ANOVA that make the former method potentially appealing as a MANOVA follow up, from a statistical perspective. Indeed, from one perspective the omnibus MANOVA test may not be required, given that an a priori ordering of the dependent variables is made by the researcher, implying a very specific set of hypotheses to be tested. Nonetheless, given the recommendations by the authors cited above to use it as a way to elucidate a significant MANOVA result, it is in this context that the current study was conducted.

As mentioned above, the SD procedure involves several steps. In the first step the researcher places the dependent variables in descending order of theoretical importance. It is crucial to note that this ordering is based on contextual issues and not on statistics. Indeed, step one should always be done prior to any data collection so that sample estimates do not affect how variables are ordered. In step two, an ANOVA is conducted in which group means on the most important response variable are compared. In step three, this variable serves as a covariate and the means of the groups on the second most important response variable are compared using ANCOVA. This stepping procedure continues for each dependent variable in turn, with response variables at a higher level of importance serving as covariates for those of lesser importance.

It has been shown that in the 2 groups case, an overall SD test statistic can be constructed that is equivalent to Hotelling's T^2 (Mudholkar & Subbaiah, 1980) allowing for an omnibus multivariate test. By including responses as covariates in subsequent analyses, correlations among the variables are accounted for in a way that they are not in the ANOVA analyses described above (Mudholkar & Subbaiah, 1975). In order to control the Type I error rate, Bock and Haggard (1968) suggested using a variation of the Bonferroni approach by dividing the overall *a* so that more important variables are given a larger share of the rejection region than less important ones, while maintaining the desired overall level of significance. It would also be possible to simply assign each test the same portion of the rejection region using Bonferroni's correction more directly (Tabachnick & Fidell, 2007).

Mudholkar and Subbaiah (1980) cited several potential advantages to using the SD procedure as a follow up to a significant MANOVA result: 1) simplicity, 2) detailed results for specific variables and groups, 3) useful with small samples and 4) results for large samples that are equivalent to the omnibus likelihood ratio test. Another potential strength of the SD is that a researcher can assign different levels of *a* to the dependent variables, reflecting their substantive importance in the investigation of which the MANOVA is a part (Subbaiah & Mudholkar, 1978; Mudholkar & Subbaiah, 1976; Bock & Haggard, 1968). Mudholkar and Subbaiah (1988) also suggest that under the SD procedure, test statistics such as the Studentized Range can be used for paired group comparisons, providing an added degree of analytic flexibility.

While there are clear benefits to the researcher using the SD technique, it does have some weaknesses as well, perhaps foremost of which is the need to order the response variables in terms of theoretical importance. It has been clearly demonstrated that the qualitative conclusions reached by researchers can differ substantially depending upon the variables' ordering (Stevens, 1972; Koslowsky & Caspy, 1991; Mudholkar & Subbaiah, 1988). Indeed, Stevens (1996) states that the SD procedure may not be appropriate when a clear a priori ordering of the response variables is not possible. Mudholkar and Subbaiah (1980) suggest that when such an ordering is not reasonable, SD can still be used as a post hoc procedure for MANOVA if the observed variables are first subjected to analysis using a data reduction technique creating linear combinations, such as Principal Components Analysis. These linear combinations could in turn be used with SD techniques, with combinations accounting for greater variance being placed higher in the sequence. Koslowsky and Caspy (1991) observed that by applying the SD to multiple sequences of the response variables a researcher could engage in meaningful data exploration and testing of multiple hypotheses. At the same time, Stevens (1996) points out that these various orderings are not independent of one another so that the actual Type I error rate across all of them is not known.

There has been some research examining the performance of the SD procedure in various conditions. Subbaiah and Mudholkar (1978) conducted a Monte Carlo simulation study in which they assessed the performance of the omnibus multivariate test statistic for the SD procedure. They simulated data sets from the multivariate normal distribution with 2 response variables, equal covariance matrices across 2 groups, each with samples of size 20. They found that the SD procedure was able to maintain the nominal Type I error rates (0.01 or 0.05) while attaining power values above 0.8 for most studied conditions. They also found that power was affected by the relative importance of the dependent variables. If they were of unequal importance and this is reflected by unequal *a* values, the power of the SD procedure was higher than when all variables were treated as equally important in terms of a. Finally, Subbaiah and Mudholkar reported that the precision of the hypothesis tests decreased for variables later in the ordering. Mudholkar and Srivastava (2000) used a robust method based on trimmed estimates to conduct a SD analysis when the data were not normally distributed and samples were of unequal size. They found that this approach had greater power than did the standard parametric ANCOVA approach.

In addition to these simulations, other researchers have reported results obtained when using the SD with real data. Stevens (1972) used both SD and DDA to compare 4 groups of subjects on 8 response variables. The two methods yielded very similar results in terms of identifying variables on which the groups differed, while the application of univariate ANOVA's produced a markedly different outcome. Stevens also demonstrated the impact on the SD analysis of different orderings of the dependent variables. Koslowsky and Caspy (1991) reported on a refinement of the SD technique in which multiple orderings of the response variables are tested in following up a significant MANOVA. If the qualitative results of these analyses differ, the researcher could conclude that there is overlap among the dependent variables. Such a result would, according to the authors, help the user gain a greater understanding of the interrelationships among the responses while also allowing for the examination of multiple hypotheses about how the groups in question might differ on the measures of interest. These authors acknowledge that when the number of dependent variables is 3 or more, the resulting Type I error rate could be somewhat inflated. Analytic results in Mudholkar and Subbaiah (1975) match the findings of their simulation study (Mudholkar & Subbaiah, 1978), demonstrating that the power of hypothesis tests for variables lower in importance (and thus tested later in the sequence) was lower than when the variables were placed higher in the sequence.

Focus of the Current Study

This study is designed to extend the work in studies reviewed above. Specifically, the goals of this research are to ascertain the performance of the SD method for following up a significant MANOVA in terms of both power and Type I error rate and to identify data specific factors influencing the performance of this approach. Given that the SD approach has been recommended in popular multivariate texts for use in the post hoc investigation of significant MANOVA results, and has been studied relatively infrequently in this role, it is hoped that this study helps to fill a gap in the literature. As has been reported above, a number of other approaches for following up a significant MANOVA have major problems that have been identified using Monte Carlo methods. For example, the STP appears to have low statistical power to investigate group differences on individual response variables, while DDA does not allow for hypothesis testing of individual dependent variables and cutoff values used with it appear to be somewhat

problematic to interpret. The two groups multivariate approach, while appearing to be reasonably effective in differentiating multivariate group means, does not easily allow for comparisons on individual variables. There has been relatively little work conducted in investigating the performance of the SD approach in terms of power and Type I error rates in the post hoc context, with most prior research focusing on its use in constructing an omnibus test statistic. It is hoped, therefore, that this study will add greater understanding in this regard.

Methodology

A Monte Carlo simulation study was used to investigate the performance of the SD follow up for MANOVA. In this case, focus was on the SD results, rather than those of MANOVA, so that results of the omnibus MANOVA were not used in the conduct of the simulation. In other words, while in actual practice the SD procedure would not be used unless a significant MANOVA were first obtained, because the focus of the current study was on the SD procedure, an assumption was made that the omnibus MANOVA test was found to be statistically significant. A number of factors were manipulated in this study with all combinations being crossed, and 1000 replications were generated for each. All simulations were conducted using SAS IML and PROC GLM. The following factors were manipulated:

Sample size. The total sample size conditions were 30, 60, 100 and 150. These values were selected in an effort to replicate sample size conditions appearing in published research using MANOVA.

Sample size ratio. Two sample size ratio conditions were simulated, including equal and unequal group sizes. In the unequal condition, the first group (group 1) had half the number of subjects as did group 2.

Number of dependent variables. Data were simulated with either 2 or 4 dependent variables. Variables earlier in the sequence were assumed to be more important than those later in the sequence and were tested as such. The lower value (2) was simulated so as to provide information about the simplest case possible, while the latter (4) was selected because it conforms to examples used in prior research (e.g., Fan & Wang, 1999).

Correlation. The dependent variables were simulated with pooled within groups correlations of 0.3, 0.5 and 0.8. All inter-variable correlations were the same in the 4 variables condition.

Effect size. A total of 8 effect size combinations, based on Cohen's *d* (Cohen, 1988), were simulated in this study to create univariate group separation. The effect sizes used were 0 (no group separation), 0.5 and 0.8. The latter values were chosen so as to correspond with Cohen's (1988) benchmarks of medium and large effects. The combinations used appear in Table 1. It is important to note that effect size values for variables 2, 3 and 4 in the 4 variable case are identical. In the case where the group variances were equal, this value was used as the denominator in the calculation of effect size, whereas in the unequal variances case, a pooled within-cell standard deviation was used.

Distribution of response variables. The dependent variables were distributed either as standard normal or with skewness of -1.5 and kurtosis of 3, with the latter distribution selected because it conforms to that used in earlier research and which has been shown to be associated with inflated Type I error rates in standard MANOVA testing (Finch, 2005). In order to maintain the desired correlation values among these variables in the non-normal condition, a method proposed by Fleishman (1977) was used to simulate the data.

Covariance matrices. The group covariance matrices were simulated to either by equal or unequal. In the latter condition, the variances for group 2 were simulated to 5 be times larger than that for group 1, which was set at a value of 1. The latter value matches that which has been demonstrated previously to lead to inflated Type I error rates for MANOVA test statistics (e.g., Finch, 2005; Sheehan-Holt, 1998).

The outcomes of interest were the Type I error rate and power for each variable in the sequence. In order to ascertain which of the manipulated factors had a significant impact on the performance of the SD follow up, ANOVA and variance components analysis were used, treating either Type I error rate or power as the dependent variable and the manipulated factors as the independent.

Table 1: Effect size combinations for simulation

Variable 1	Variable 2 (3 and 4)
0.5	0
0.8	0
0	0.5
0	0.8
0.5	0.5
0.8	0.8
0.8	0.5
0.5	0.8

Results

Type I error rate

The results of the ANOVA/variance components analyses for the 2 and 4 variable cases yielded similar results in terms of identifying factors that influenced the Type I error rate of the SD procedure. For this reason, results of the ANOVA are only reported for the 2 variable case. Results for the second variable in the two groups case are nearly identical to those for variables 2, 3 and 4 in the 4 variable case. Specifically, for the Type I error rate of the first response variable (variable 1), the 2-way interaction of equality of the covariance matrices and the distribution of the variables was the highest order significant interaction, and accounted for 90.9% of the observed variability. The 3-way interaction of the correlation between the response variables by equality of the covariance matrices by variable distribution was the highest order significant term for the second variable (variable 2), accounting for 46% of the variation in Type I error rates. The 2-way interaction of correlation and response variable distribution was also significant for variable 2 and accounted for an additional 20% of variation in Type I error rate. Other main effects and interactions were also statistically significant for the Type I error rates of both variables, but none accounted for more than 10% of the variation, and will therefore not be considered further.

Table 2 contains the Type I error rates for the 2 and 4 variable cases by the inter-variable correlation, covariance status (equal or unequal) and distribution of the response variables. It appears that the Type I error rate for the first variable in the set is near the nominal level of 0.05, except when the assumptions of normality and equality of covariance matrices underlying the ANCOVA have both been violated. The Type I error rates for variables 2 through 4 were found to be inflated for most of the conditions displayed in Table 2. The Type I error inflation for these variables also increased somewhat with increasing correlation among the variables. Although all Type I error conditions for the later variables in the sequence were inflated, the severity of this inflation was less when the assumptions of normality and equality of covariance matrices were both met.

Power

The ANOVA and variance components analyses for the power rates indicated that for variable 1, the statistically significant interaction between covariance matrix equality/inequality and distribution of the variables accounted for 78.5% of the variability in power. No other term in the model was both statistically significant and accounted for more than 10% of the variance in power for variable 1. Three terms were found to be statistically significant and account for more than 10% each of the variation in power for variable 2 (3 and 4). The highest order interaction, accounting for 25% of power variation, was covariance matrix equality/inequality by correlation by distribution, while effect size (15.5%) and sample size (12.1%) were the two statistically significant main effects that accounted for more than 10% of variation in power. As with the Type I error rates, the same factors contributed significantly to observed power in the 2 and 4 variable cases.

Table 3 displays the power rates by covariance matrix equality/inequality, correlation and distribution for all of the variables. One outcome to note is that across these conditions power rates generally diminish through the sequence from variable 1 to 4. Power for the first variable in the set was higher in the normal distribution condition when the groups' covariance matrices were also equal than when the data were normal but the variances were unequal. However, when the covariance matrices were unequal, the opposite pattern can be seen, where variable 1 power was greater in the skewed condition. It

Table 2: Type I error rate by correlation among dependent variables, covariance equality/inequality and response distribution: 2 and 4 variables

2 response variables				
Correlation	Covariance	Distribution	V1	V2
0.3	Equal	Normal	0.051	0.141
		Skewed	0.047	0.252
	Unequal	Normal	0.036	0.063
		Skewed	0.848	0.674
0.5	Equal	Normal	0.047	0.142
		Skewed	0.044	0.259
	Unequal	Normal	0.038	0.132
		Skewed	0.852	0.784
0.8	Equal	Normal	0.053	0.234
		Skewed	0.046	0.317
	Unequal	Normal	0.034	0.488
		Skewed	0.850	0.827

4 response variables						
Correlation	Covariance	Distribution	V1	V2	V3	V4
0.3	Equal	Normal	0.050	0.143	0.138	0.107
		Skewed	0.047	0.251	0.212	0.181
	Unequal	Normal	0.050	0.144	0.144	0.140
		Skewed	0.849	0.674	0.540	0.512
0.5	Equal	Normal	0.053	0.135	0.125	0.150
		Skewed	0.044	0.266	0.267	0.249
	Unequal	Normal	0.036	0.137	0.133	0.151
		Skewed	0.852	0.690	0.652	0.640
0.8	Equal	Normal	0.051	0.236	0.228	0.208
		Skewed	0.049	0.366	0.450	0.362
	Unequal	Normal	0.034	0.490	0.324	0.287
		Skewed	0.849	0.870	0.831	0.830

is important to keep in mind that it was in this combination of conditions (unequal covariance matrices and skewed data) that the Type I error rate for variable 1 was elevated. For this reason, these high power rates in the skewed/nonnormal condition are not particularly meaningful. With respect to variables 2, 3 and 4, a pattern similar to that for variable 1 was evident, where power was higher for the normal (versus skewed) distribution when the covariances were also equal, while the opposite pattern by distribution was evident when the covariance matrices were unequal. Again, given the aforementioned Type I error inflation when neither assumption was met, it is not meaningful (nor particularly surprising) that power in this case was also higher.

Table 4 includes power by effect size combination. Please note that the category "50" in the table refers to the condition where the first variable has an effect size difference of 0.5 between the groups for the first response variable and no difference for variable 2 (as well as 3 and 4 where applicable). In like fashion, "05" refers to the condition where the first variable has no group difference but variable 2 (3 and 4) is characterized by an effect size difference of 0.5. The remaining combinations should be interpreted similarly. Results in this table for variable 1 show that a greater effect size, reflecting larger group separation, was associated with higher power. Furthermore, the power for variable 1 was unaffected by the effect size of the accompanying variable(s). Of course, this would be expected given that it is the first variable in the sequence, and thus tested independently of the others.

Results in Table 4 demonstrate that, as with variable 1, greater effect size was associated with higher power for variables 2 through 4. However, an important outcome for these latter variables was that the effect size of the first variable had an impact on their power. For example, in the two variables condition, the power for the second variable in the 05 case (where there was not a group difference for the first response and a difference of 0.5 for the second response) was 0.365, while when the first variable was characterized by a large effect size difference and the second by this same moderate effect (85), the

Table 3: Power by correlation among dependent variables, covariance equality/inequality and response distribution: 2 and 4 variables

2 response variables				
Correlation	Covariance	Distribution	V1	V2
0.3	Equal	Normal	0.721	0.513
		Skewed	0.193	0.122
	Unequal	Normal	0.359	0.243
		Skewed	0.872	0.693
0.5	Equal	Normal	0.718	0.454
		Skewed	0.194	0.176
	Unequal	Normal	0.357	0.228
		Skewed	0.873	0.136
0.8	Equal	Normal	0.717	0.456
		Skewed	0.194	0.278
	Unequal	Normal	0.355	0.297
		Skewed	0.872	0.070

4 response variables						
Correlation	Covariance	Distribution	V1	V2	V3	V4
0.3	Equal	Normal	0.687	0.360	0.400	0.301
		Skewed	0.176	0.097	0.096	0.073
	Unequal	Normal	0.690	0.361	0.397	0.296
		Skewed	0.873	0.682	0.522	0.398
0.5	Equal	Normal	0.688	0.352	0.323	0.218
		Skewed	0.175	0.173	0.076	0.057
	Unequal	Normal	0.336	0.150	0.144	0.099
		Skewed	0.869	0.593	0.544	0.531
0.8	Equal	Normal	0.691	0.464	0.315	0.207
		Skewed	0.175	0.312	0.079	0.052
	Unequal	Normal	0.334	0.254	0.157	0.103
		Skewed	0.875	0.535	0.522	0.521

Table 4: Power by effect size: 2 and 4 variables

2 response variables				
Effect size combination	V1		V2	
50 / 05*	0.424		0.365	
80 / 08*	0.610		0.582	
85	0.610		0.188	
58	0.426		0.449	
88	0.611		0.235	
55	0.424		0.158	
4 response variables				
Effect size combination	V1	V2	V3	V4
50 / 05*	0.454	0.395	0.237	0.163
80 / 08*	0.641	0.613	0.365	0.255
85	0.643	0.199	0.115	0.092
58	0.454	0.479	0.267	0.183
88	0.643	0.263	0.177	0.132
55	0.452	0.178	0.126	0.099

*The effect size combination to the left of / is for variable 1, while to the right of / is the corresponding combination for variable 2.

Table 5: Power by sample size: 2 and 4 variables

2 response variables				
Sample size	V1		V2	
30	0.296		0.158	
60	0.509		0.263	
100	0.626		0.359	
150	0.711		0.441	
4 response variables				
Sample size	V1	V2	V3	V4
30	0.307	0.140	0.102	0.080
60	0.524	0.235	0.173	0.121
100	0.640	0.330	0.255	0.178
150	0.719	0.406	0.327	0.239

power for variable 2 was 0.188. In other words, the power for detecting mean differences for variable 2 decreased between the two examples, though the degree of group separation in the two cases was identical. This result was present across all effect size combinations for all three latter variables in the sequence. As noted above, power declined from variable 2 through variable 4 in the testing sequence. This outcome was true even though the variables were characterized by comparable effect sizes in the 4 variables condition. In other words, even though in the population group separation was equivalent for variables 2, 3 and 4, power was higher for those variables tested earlier in the sequence.

Table 5 contains power rates by sample size. In interpreting all of these power results, it is important to keep in mind that the Type I error rates presented in Table 2 were elevated above the nominal error rate of 0.05 for many of the conditions studied here. Power for all variables in both the 2 and 4 variable conditions increased concomitantly with increases in sample size. As was evident from results in Table 4, power was lower for variables tested later in the sequence than those tested earlier.

Discussion

The purpose of this study was to explore the effectiveness of the Roy-Bargmann stepdown procedure as a follow up to a significant MANOVA as has been recommended in some popular multivariate statistics texts in the social sciences (Tabachnick & Fidell, 2007; Stevens, 1996). Using this method, a researcher would conclude that dependent variables for which significant results on the individual ANCOVA's were found could be identified as "important" contributors to the group separation identified by the MANOVA (Mudholkar and Subbaiah, 1980). The results of this study indicate that when both the assumptions of normality and equality of covariance matrices are not met, the Type I error rate for the first variable in the sequence was inflated, while in all other cases it was near the nominal 0.05 level. This result should serve to encourage researchers using the SD as a post hoc to a significant MANOVA to assess the viability of both the normality and equality of covariance matrices assumptions prior to making use of the technique. However, the Type I error rates of subsequent variables in the sequence were inflated in most cases studied here, including when the assumptions underlying ANCOVA are met. This widespread Type I error inflation for the latter variables in the testing sequence must call into question the findings for power, though they have been presented in order to fully inform the reader as to the results of the study. In general, power rates declined further into the testing sequence so that even with the same effect size in the population, power was lower for testing variable 2 than for variable 1 and for variable 3 than for variable 2, etc. Furthermore, the power for later variables in the sequence was influenced by the effect size of the first variable, such that larger group differences on variable 1 were associated with lower power on variables 2, 3 and 4, regardless of the effect size characterizing these latter variables.

Implications for practice

It should be remembered that this study was conducted under the presumption that a researcher would elect to use the SD approach prior to the conduct of analysis. Thus, by implication a reasonable ordering of the variables, based on contextual issues germane to the area being studied and not statistical concerns, is possible. In cases where this ordering is not reasonable, the SD procedure would not be an

appropriate approach, and some other method would clearly seem to be better as a follow-up for MANOVA.

The results of this research must, to some degree, call into question the use of the Roy-Bargmann stepdown procedure as a follow-up to a significant MANOVA, despite recommendations for its use (e.g., Tabachnick & Fidell, 2007; Stevens, 1996). Prior research (e.g., Kabe, 1984; Mudholkar & Subbaiah, 1980) has shown that this approach works well in many conditions in terms of constructing an omnibus test of significance in the multivariate case. However, the current study does not support the use of this technique in trying to identify specific variables for which the groups might differ, other than the first one. When groups didn't differ on the latter variables in the testing sequence, the Type I error rate was generally inflated above the nominal rate, while when they did differ, power tended to be lower than for variable 1. Furthermore, in cases where the first variable in the sequence was characterized by a large effect size (0.8), power for the latter variables was even more diminished.

For these reasons, it does not seem prudent to use this technique to ascertain which variables might be significantly different between groups. This does not call into question the potential utility of this procedure for an omnibus test in the MANOVA context, as mentioned previously. In that case, the hypothesis being tested is very different from those addressed by the method as studied here. In addition, given the inflated Type I error rates for the tests of all but variable 1, practitioners who do choose to use this approach in the manner outlined here should be very wary of significant results for all but the first variable in the testing sequence, particularly in (though not limited to) cases when the assumptions of normality and homogeneity of variances are not met.

Limitations and directions for future research

There are some methodological limitations to this study that should be addressed in future research efforts. Perhaps chief among these is the somewhat limited set of effect size values used. The 6 combinations reported here were selected in order to allow for an investigation of Type I error rate and power under a variety of conditions while maintaining a manageable study size. These values selected corresponded to the benchmarks of medium and large effects laid out by Cohen (1988). It is recognized, however, that more work is needed with a greater variety of patterns in order to gain a more complete understanding of the performance of the Roy-Bargmann approach to post hoc testing for a significant MANOVA. Such a future effort should include both a different set of effect size values, particularly those smaller than 0.5, as well as a more complex pattern of values in the 4 variables condition. In the current study, variables 2, 3 and 4 all had the same effect size value for all conditions, which may not be representative of the wide array of possibilities seen in real data. At the same time, given that this study is one of the first Monte Carlo efforts to examine the performance of the Roy-Bargmann method in this MANOVA post hoc context, it was felt that in order to maintain a manageable study design and ease interpretation, this somewhat limited set of combinations was advisable. Furthermore, it is not clear that major differences in results would be expected from those presented here.

A second recommendation for future research would be to include more group conditions. As stated above, it was felt that since this study was examining this approach in a fairly different way from previous research, a simpler design was preferable. However, in order to broaden its generalizability, future efforts should include more groups.

References

Bird, K.D. & Hadzi-Pavlovic, D. (1983). Simultaneous test procedures and the choice of a test statistic in MANOVA. *Psychological Bulletin, 93*, 167-178.

Bock, R.D. & Haggard, E. (1968). The use of Multivariate Analysis of Variance in research. In D.K. Whitla (Ed.), *Handbook of measurement assessment in behavioral sciences*. Reading, MA: Addison Wesley.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, Hillsdale, NJ: Lawrence Erlbaum.

Elliott, R. (1993). *Contrast selection in post hoc multivariate analysis*. Unpublished Doctoral Dissertation, Ohio University-Chillicothe.

Enders, C.K. (2003). Performing multivariate group comparisons following a statistically significant MANOVA. *Measurement and Evaluation in Counseling and Development, 36*, 40-56.

Fan & Wang. (1999). Comparing linear discriminant function with logistic regression for the two-group classification problem. *The Journal of Experimental Education*, 67(3), 265-286.

Fleishman, A.I. (1977). A method for simulating non-normal distributions. *Psychometrika, 43*, 521-532.

Finch, W.H. (2005). Comparison of the performance of nonparametric and parametric MANOVA test statistics when assumptions are violated. *Methodology*, 1, 27-38.

Gabriel, K. (1968). Simultaneous test procedures in MANOVA. *Biometrika, 55*, 489-504.
Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, 360-378.
Huberty, C.J. (1994). *Applied Discriminant Analysis.* New York: John Wiley and Sons, Inc.
Kabe, D.G. (1984). On the maximal invariance of MANOVA stepdown procedure statistics. *Communications in Statistics: Theory and Methods, 13*, 2571-2581.
Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donohue, B., Kowalchuk, R.K., Lowman, L.L., Petosky, M.D., Keselman, J.C., & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA and ANCOVA analyses. *Review of Educational Research, 68*, 350-368.
Kirk, R.E. (1995). *Experimental Design*. Pacific Grove, CA: Brooks/Cole Publishing Company.
Kowlowsky, M. & Caspy, T. (1991). Stepdown analysis of variance: A refinement. *Journal of Organizational Behavior, 12*, 555-559.
Marden, J.I. & Perlman, M.D. (1990). On the inadmissibility of step-down procedures for the Hotelling T^2 problem. *The Annals of Statistics, 18*, 172-190.
Maxwell, S.E. (1992). Recent developments in MANOVA applications. In *Advances in Social Science Methodology: Volume 2* (pp.137-168). Greenwich, CT: JAI Press.
Mudholkar, G.S., Davidson, M.L. & Subbaiah, P. (1974). Extended linear hypotheses and simultaneous tests in Multivariate Analysis of Variance. *Biometrika, 61*, 467-477.
Mudholkar, G.S. & Srivastava, D.K. (2000). A class of robust stepwise alternatives to Hotelling's T^2 tests. *Journal of Applied Statistics, 27*, 5999-619.
Mudholkar, G.S. & Subbaiah, P. (1975). A note on MANOVA multiple comparisons based upon step-down procedure. *The Indian Journal of Statistics, 37*, 300-307.
Mudholkar, G.S. & Subbaiah, P. (1980). A review of step-down procedures for Multivariate Analysis of Variance. In R.P. Gupta, (Ed.), *Multivariate Statistical Analysis (pp.161-178)*. Amsterdam: North-Holland.
Mudholkar, G.S. & Subbaiah, P. (1988). On a fisherian detour of the step-down procedure for MANOVA. *Communications in Statistics: Theory and Methods, 17*, 599-611.
Olson, C.L. (1974). Comparative robustness of six tests in Multivariate Analysis of Variance. *Journal of the American Statistical Association, 69*, 894-908.
Rencher, A.C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician, 46*, 217-225.
Roy, J. (1958). Step-down procedure in multivariate analysis. *Annals of Mathematical Statistics, 29*, 1177-87.
Roy, S.N. & Bose, R.C (1953). Simultaneous confidence interval estimation. *Annals of Mathematical Statistics, 24*, 513-536.
Roy, S.N. & Bargmann, R.E. (1958). Tests of multiple independence and the associated confidence bounds. *The Annals of Mathematical Statistics, 29*, 491-503.
Schneider, M. K. (2004). 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. *Multiple Linear Regression Viewpoints, 30*, 8-18.
Sheehan-Holt, J.K. (1998). MANOVA simultaneous test procedures: The power and robustness of restricted multivariate contrasts. *Educational and Psychological Measurement, 58*, 861-881.
Sheehan, J.K. (1995). A comparison of the Type I error and power of selected MANOVA simultaneous test procedures. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA, April 18-22.
Stevens, J.P. (1972). Four methods of analyzing between variation for the K-group MANOVA problem. *Multivariate Behavioral Research, 7*, 499-522.
Stevens, J.P. (1996). *Applied Multivariate Statistics for the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum and Associates, Publishers.
Subbaiah, P. & Mudholkar, G.S. (1978). A comparison of two tests for the significance of a mean vector. *Journal of the American Statistical Association, 73*, 414-418.
Tabachnick, B.G. & Fidell, L.S. (2007). *Using Multivariate Statistics*. Boston: Pearson.

Send correspondence to: W. Holmes Finch
Ball State University
Email: whfinch@bsu.edu