Multiple Linear Regression Viewpoints

A Publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model

MLRV

Volume 34 • Number 1 • Fall 2008

Table of Contents

Multiple Linear Regression Viewpoints

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: General Linear Model through the **University of Alabama at Birmingham**.

Subscription and SIG membership information can be obtained from:
Cynthia Campbell, MLR:GLM/SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854.
ccampbell@niu.edu

MLRV abstracts appear in CIJE, the ERIC system, and microform copies are available from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106. *MLRV* is listed in the *EBSCO Librarians Handbook*.

Seemingly Unrelated Regression (SUR) Models as a Solution to Path Analytic Models with Correlated Errors

T. Mark Beasley
University of Alabama at Birmingham

Multivariate regression requires the design matrix for each of *p* dependent variables to be the same in form. Zellner (1962) formulated Seemingly Unrelated Regression (SUR) models as *p* correlated regression equations. SUR models allow each of the *p* dependent variables to have a different design matrix with some of the predictor variables being the same. Of particular relevance to path analysis, SUR models allow for a variable to be both in the **Y** and **X** matrices. SUR models are a flexible analytic strategy and are underutilized in educational research.

Standard multivariate regression requires that each of *p* dependent variables has exactly the same design matrix such that:

$$\mathbf{Y}_{(N\text{x}p)} = \mathbf{X}_{(N\text{x}k)}\,\mathbf{B}_{(k\text{x}p)} + \boldsymbol{\varepsilon}_{(N\text{x}p)}\,, \tag{1}$$

where **Y** is a matrix of *p* dependent variables, **X** is a *k*-dimensional design matrix, and **ε** is an error matrix, which is assumed to be distributed as $\mathcal{N}_{(N\text{x}p)}(\mathbf{0},\Sigma\otimes\mathbf{I}_N)$. Multivariate regression theory using ordinary least squares (OLS) assumes that all of the **B** coefficients in the model are unknown and to be estimated from the data as:

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})\,. \tag{2}$$

Multivariate Regression and Multiple Univariate Regression

Multivariate regression is not used often in behavioral research. One reason is that the matrix algebra underlying parameter estimation (2) is a column solution. Thus, whether one uses multivariate regression or *p* separate univariate regression, the regression coefficients will be the same. The differences between univariate and multivariate regression are the types of hypotheses that can be tested and the standard errors for these secondary parameters. Suppose one were to regress $p = 3$ dependent variables (y) on to $k = 2$ predictor variables (x). The omnibus null hypothesis would be that the regression coefficients for both x_1 and x_2 on all three y variables equals zero, a multivariate test with 6 degrees-of-freedom (*df*). Another hypothesis of potential interest would that x_2 has no unique relationship to any of the three y variables after controlling for x_1, a multivariate test with $df = 3$. It is important to note that if one were to construct a "multivariate" test that reduced down to only one of the y variables, then the results will be the same as the univariate regression, which is another reason multivariate regression is not popular.

Zellner (1962) formulated the Seemingly Unrelated Regression (SUR) model as *p* correlated regression equations. The *p* regression equations are "seemingly unrelated" because taken separately the error terms would follow standard linear OLS model form. Calculating *p* separate standard OLS solutions ignores any correlation among the errors across equations; however, because the dependent variables are correlated and the design matrices may contain some of the same variables there may be "contemporaneous" correlation among the errors across the *p* equations. Thus, SUR models are often applied when there may be several equations, which appear to be unrelated; however, they may be related by the fact that: 1) some coefficients are assumed to be the same or zero; 2) the disturbances are correlated across equations; and/or 3) a subset of right hand side variables are the same. This third condition is of particular interest because it allows each of the *p* dependent variables to have a different design matrix with some of the predictor variables being the same. SUR models allow for a variable to be both in the **Y** and **X** matrices, which has particular relevance to path analysis. SUR models are an underused multivariate technique. Using SUR models to solve path analytic models will be explicated.

SUR Model

The SUR model is a generalization of multivariate regression using a vectorized parameter model. The **Y** matrix is vectorized by vertical concatenation, **y**v. The design matrix, **D**, is formed as a block diagonal with the j^{th} design matrix, **X***j*, on the j^{th} diagonal block of the matrix. The model is then expressed as:

$$\text{E}[\mathbf{Y}_{(N\text{x}p)}] = \{\mathbf{X1}_{(N\text{x}m_1)}\,\beta\mathbf{1}_{(m_1\text{x}1)}\,,\ \mathbf{X2}_{(N\text{x}m_2)}\,\beta\mathbf{2}_{(m_2\text{x}1)}\,,\ \mathbf{X}\boldsymbol{j}_{(N\text{x}m_j)}\,\beta\boldsymbol{j}_{(m_j\text{x}1)}\,,\ \mathbf{X}\boldsymbol{p}_{(N\text{x}m_p)}\,\beta\boldsymbol{p}_{(m_p\text{x}1)}\}\ ; \tag{3}$$

where *mj* is the number of parameters estimated (columns) by the j^{th} design matrix, $\mathbf{X}_j$.

To illustrate in matrix notation, the SUR model is laid out as:

$$\underset{(Np\text{x}1)}{\mathrm{E}(\mathbf{y}_v)} = \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \dots \\ \hat{\mathbf{y}}_j \\ \dots \\ \hat{\mathbf{y}}_p \end{bmatrix}_{(Np\text{x}1)} \begin{matrix} (N\text{x}1) \\ (N\text{x}1) \\ \\ (N\text{x}1) \\ \\ (N\text{x}1) \end{matrix} = \underset{(Np\text{x}M)}{\begin{bmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} & \dots & \mathbf{0} \\ (N\text{x}m_1) & \mathbf{X}_2 & \dots & \mathbf{0} & \dots & \mathbf{0} \\ & (N\text{x}m_2) & \dots & \mathbf{0} & \dots & \mathbf{0} \\ & & & \mathbf{X}_j & \dots & \mathbf{0} \\ & & (sym) & (N\text{x}m_j) & & \mathbf{X}_p \\ & & & & & (N\text{x}m_p) \end{bmatrix}} \underset{(M\text{x}1)}{\begin{bmatrix} \beta_1(m_1\text{x}1) \\ \beta_2(m_2\text{x}1) \\ \\ \beta_j(m_j\text{x}1) \\ \\ \beta_p(m_p\text{x}1) \end{bmatrix}} ; \quad (4)$$

where M is the total number of parameters estimated over the p models, $M = \sum_{j=1}^{p} m_j$.

Estimators for the SUR Model

One approach to solving the parameter estimates is:

$$\underset{(M\text{x}Np)}{[\mathbf{D}'} \; \underset{(Np\text{x}Np)}{\mathbf{Q}^{-1}} \; \underset{(Np\text{x}M)}{\mathbf{D}]^{-1}} \; \underset{(M\text{x}Np)}{[\mathbf{D}'} \; \underset{(Np\text{x}Np)}{\mathbf{Q}^{-1}} \; \underset{(Np\text{x}1)}{\mathbf{y}_v]} \quad (5)$$

Q is weight matrix based on the residual covariance matrix of the **Y** variables and is formed as:

$$\underset{(Np\text{x}Np)}{\mathbf{Q}} = \underset{(p\text{x}p)}{\hat{\boldsymbol{\Sigma}}} \otimes \mathbf{I}_N . \quad (6)$$

To elucidate, the residual covariance matrix could be computed by regressing each of the p dependent variables on to its design matrix and obtaining the residuals. The j^{th} diagonal element of $\hat{\boldsymbol{\Sigma}}$ is computed by calculating the Sum of Squares for the j^{th} residual. The ij^{th} off-diagonal element is computed by taking the cross-product of the i^{th} and j^{th} residuals. These values are then divided by an estimate for the degrees-of-freedom for each element. Using matrix notation, the ij^{th} element of $\hat{\boldsymbol{\Sigma}}$ is calculated as:

$$\hat{\sigma}_{ij} = \frac{1}{(N - df^*)} \mathbf{y}'_i [\mathbf{I}_N - \mathbf{H}_i][\mathbf{I}_N - \mathbf{H}_j] \mathbf{y}_j ; \quad (7)$$

where $\mathbf{H}j = \mathbf{X}_j(\mathbf{X}'_j \mathbf{X}_j)^{-1}\mathbf{X}'_j$ is the hat matrix for the j^{th} design matrix. Although there are several approaches for defining the degrees of freedom, the most common approach is to define *df** as the average of the numerator degrees-of-freedom (*df*) for the i^{th} and j^{th} models. Thus, this SUR estimator, sometimes referred to as Zellner's two-stage Aitken estimator, is an application of generalized least squares (GLS). In fact, because the residual covariance matrix is unknown and must be estimated from the data, this application is often called feasible generalized least squares (FGLS; see Timm, 2002). It should be noted that if $\mathbf{Q}^{-1}$ is removed from equation (5), or is defined as an identity matrix ($\mathbf{Q}^{-1} \equiv \mathbf{I}$), then the results will be the same as p separate univariate regression models. To develop robust standard errors or more precise estimates of $\boldsymbol{B}$, Zellner (1962) also proposed iterating the FGLS solution (IFGLS), which has the same asymptotic properties as the FGLS (Kmenta & Gilbert, 1968). To obtain maximum likelihood (ML) estimators of $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$, Kmenta and Gilbert (1968) employed an iterative procedure to solve the likelihood equation:

$$L(\boldsymbol{B}, \boldsymbol{\Sigma} \mid \mathbf{y}) = (2\pi)^{-Np/2} \mid \boldsymbol{\Sigma} \mid^{-N/2} e^{\mathrm{tr}[(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I})(\mathbf{y} - \mathbf{D}\boldsymbol{B})(\mathbf{y} - \mathbf{D}\boldsymbol{B})']} \quad (8)$$

Park (1993) showed that the ML and IFGLS estimators are mathematically equivalent. Kmenta and Gilbert (1968) found that the ML (IFGLS) and FGLS estimators gave similar results; however, FGLS is favored in small samples. Because the FGLS estimator is always unbiased and requires the least computation burden, it is recommended in most applications of the SUR model with small samples.

SUR Model Approach to Path Analysis

To demonstrate how a SUR model can be used to solve a path analysis problem, suppose the path model in Figure 1. The "terminal" endogenous variable is y_1, which is directly influenced by y_2, y_3, and x_2. One exogenous variable, x_2 also has indirect effects on y_1 through y_2 and y_3. The exogenous variable, x_1, has an indirect effect on y_1 though y_2. The exogenous variable, x_3, has an indirect effect on y_1 though y_3. The path diagram also models correlation among the errors of the endogenous variables. Assuming standardized variables so that all intercepts will be zero, the correctly specified regression models would be:

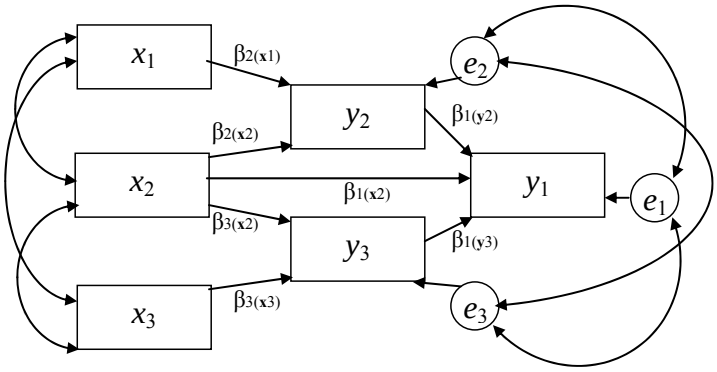

Figure 1. Hypothetical Path Model

$$
\begin{aligned}
\hat{\mathbf{y}}_1 &= \beta_{1(\mathbf{y}2)}\,\mathbf{y}_2 + \beta_{1(\mathbf{y}3)}\,\mathbf{y}_3 + \mathbf{0}\,\mathbf{X}_1 + \beta_{1(\mathbf{x}2)}\,\mathbf{X}_2 + \mathbf{0}\,\mathbf{X}_3 \\
\hat{\mathbf{y}}_2 &= \beta_{2(\mathbf{x}1)}\,\mathbf{X}_1 + \beta_{2(\mathbf{x}2)}\,\mathbf{X}_2 + \mathbf{0}\,\mathbf{X}_3 \\
\hat{\mathbf{y}}_3 &= \mathbf{0}\,\mathbf{X}_1 + \beta_{3(\mathbf{x}2)}\,\mathbf{X}_2 + \beta_{3(\mathbf{x}3)}\,\mathbf{X}_3
\end{aligned}
\tag{9}
$$

The first subscript refers to the dependent variable (y) and the second subscript in parentheses refers to the predictor variable. For example, $\beta_{1(\mathbf{y}3)}$ refers to the regression coefficient (path) of y_3 to y_1. Because the dependent variables and their error terms are correlated and the design matrices contain some of the same variables there is "contemporaneous" correlation among the errors across the p equations. However, the standard OLS solutions will ignore any correlation among the errors across these three equations.

Appendix B shows SAS/IML code for generating data for the path model in Figure 1. The sample size was set at N = 5000 so that asymptotical properties could be observed. The correlations among the exogenous **X** variables were set at $r_{\mathbf{x}12}$ = 0.30, $r_{\mathbf{x}13}$ = 0.25., and $r_{\mathbf{x}23}$ = 0.15. Table 1 displays the other preset coefficients.

Solving Parameter Estimates for SUR Models

The correctly specified SUR model for this path analytic problem would be laid out as such:

$$
\underset{(3N\text{x}1)}{\mathrm{E}(\mathbf{y}_v)} = \underset{(3N\text{x}1)}{\begin{bmatrix} \hat{\mathbf{y}}_{11} \\ \hat{\mathbf{y}}_{12} \\ \vdots \\ \hat{\mathbf{y}}_{1N} \\ \hat{\mathbf{y}}_{21} \\ \hat{\mathbf{y}}_{22} \\ \vdots \\ \hat{\mathbf{y}}_{2N} \\ \hat{\mathbf{y}}_{31} \\ \hat{\mathbf{y}}_{32} \\ \vdots \\ \hat{\mathbf{y}}_{3N} \end{bmatrix}} = \underset{(3N\text{x}7)}{\begin{bmatrix} \mathbf{y}_{21} & \mathbf{y}_{21} & \mathbf{x}_{21} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{y}_{22} & \mathbf{y}_{22} & \mathbf{x}_{22} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & & & & \\ \mathbf{y}_{2N} & \mathbf{y}_{21} & \mathbf{x}_{21} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ & (N\text{x}3) & & \mathbf{x}_{11} & \mathbf{x}_{21} & \mathbf{0} & \mathbf{0} \\ & & & \mathbf{x}_{12} & \mathbf{x}_{22} & \mathbf{0} & \mathbf{0} \\ & & & \vdots & \vdots & & \\ & & & \mathbf{x}_{1N} & \mathbf{x}_{2N} & \mathbf{0} & \mathbf{0} \\ & (sym) & & (N\text{x}2) & & \mathbf{x}_{21} & \mathbf{x}_{31} \\ & & & & & \mathbf{x}_{22} & \mathbf{x}_{32} \\ & & & & & \vdots & \vdots \\ & & & & (N\text{x}2) & \mathbf{x}_{2N} & \mathbf{x}_{3N} \end{bmatrix}} \underset{(7\text{x}1)}{\begin{bmatrix} \beta_{1(\mathbf{y}2)} \\ \beta_{1(\mathbf{y}3)} \\ \beta_{1(\mathbf{x}2)} \\ \beta_{2(\mathbf{x}1)} \\ \beta_{2(\mathbf{x}2)} \\ \beta_{3(\mathbf{x}2)} \\ \beta_{3(\mathbf{x}3)} \end{bmatrix}}
\tag{10}
$$

Setting this path analysis model up as a SUR model allows for the simultaneous solution of the coefficients in closed form and will produce estimates of the standard errors that take the contemporaneous correlations into account.

Appendix C shows code for the SYSLIN, CALIS, and MIXED modules of SAS. In the PROC SYSLIN code the FIML option produces the Full Information Maximum Likelihood estimates. Other estimation methods include the SUR option, which produces the FGLS estimates, and the ITSUR (ITterative SUR) option, which produces the IFGLS estimates.

Table 1. Parameter Estimates from SAS PROC SYSLIN, CALIS, and MIXED.

Coefficients for:	y_1			y_2		y_3		Correlations for Errors		
Parameter	$\beta_{1(y2)}$	$\beta_{1(y3)}$	$\beta_{1(X2)}$	$\beta_{2(X1)}$	$\beta_{2(X2)}$	$\beta_{3(X2)}$	$\beta_{3(X3)}$	r_{e12}	r_{e13}	r_{e23}
Values	**0.25**	**0.35**	**0.20**	**0.35**	**0.20**	**0.40**	**0.25**	**0.10**	**0.20**	**0.10**
SYSLIN (FIML)	0.2542 (0.0299)	0.3338 (0.0421)	0.2106 (0.0215)	0.3600 (0.0131)	0.1972 (0.0132)	0.4041 (0.0123)	0.2503 (0.0123)	NA	NA	NA
CALIS (ML)	0.2542 (0.0299)	0.3338 (0.0421)	0.2106 (0.0215)	0.3600 (0.0131)	0.1972 (0.0132)	0.4041 (0.0123)	0.2503 (0.0123)	0.0989	0.2135	0.1048
MIXED (ML)	0.2542 (0.0106)	0.3338 (0.0112)	0.2106 (0.0116)	0.3600 (0.0131)	0.1972 (0.0132)	0.4041 (0.0123)	0.2503 (0.0121)	0.0989	0.2135	0.1048
SYSLIN (SUR)	0.2980 (0.0106)	0.4803 (0.0112)	0.1324 (0.0114)	0.3604 (0.0131)	0.1971 (0.0132)	0.4040 (0.0123)	0.2511 (0.0123)	0.0146	0.0176	0.1047
SYSLIN (ITSUR)	0.2521 (0.0106)	0.3495 (0.0112)	0.2043 (0.0116)	0.3600 (0.0131)	0.1972 (0.0132)	0.4040 (0.0123)	0.2511 (0.0121)	0.0998	0.1943	0.1048

Note: Standard Errors are in parentheses under the parameter estimates.

Another approach to solving the parameter estimates is to set the equations up as a multivariate (or SUR in this case) linear mixed model (LMM) and use SAS PROC MIXED. However, multivariate LMMs have received scant treatment in the literature. Reinsel (1984) derived closed-form estimates with completely observed data and balanced designs. More recently, Shah, Laird, and Schoenfeld (1997) extended the EM-type algorithm of Laird and Ware (1982) to a bivariate ($p = 2$) setting. In econometric terminology, their model is analogous to SUR. Schafer and Yucel (2002) note that the added generality of the SUR model comes at a high cost, making the resulting algorithms impractical for more than a few response variables. Thus, it may be possible to recast the multivariate model as a univariate one by stacking the columns of y_j and applying SAS PROC MIXED with a user-specified covariance structure (see Appendix B for the code to stack the data). In most applications, however, this approach quickly becomes impractical. Examples for only $p = 2$ response variables with complete data (Shah et al., 1997) and incomplete data (Verbeke & Molenberghs, 2000) require complicated SAS macros. As the number of variables and number of individuals per cluster grows, the dimension of the response vector increases rapidly, and usage of SAS PROC MIXED can become practically impossible.

Fortunately, Park (1993) showed that the ML and IFGLS estimators are mathematically equivalent. As can be seen in Table 1 the estimates from PROC MIXED with an ML estimator and PROC CALIS with an ML estimator produce identical parameter estimates but slightly different standard errors. The results from PROC SYSLIN with the ITSUR option (IFGLS estimator) are virtually identical to those from PROC MIXED. PROC CALIS with an ML estimator and PROC SYSLIN with the FIML option produce identical parameter estimates and standard errors, but PROC SYSLIN does not report the correlation among the regression equations (error terms for the *y* variables). The SUR (FGLS) option gives similar results but the solution has not been iterated as in the ITSUR (IFGLS) option. A full-scale simulation study would be necessary to determine which approach would provide the most accurate and valid results. A researcher interested in conducting a simulation study could compare the bias in the coefficients and standard errors of the correctly specified regression (9) and SUR (10) models and the results from structural equation modeling software (e.g. SAS PROC CALIS). One could also assess power and Type I error of correctly specified and misspecified models. For example, on could analyze a model that incorrectly assumes a direct path from x_1 to y_1 and then investigate the Type I error rates produced by the different analytic approaches. Furthermore, one could compare the statistical properties of different estimation procedures under any of these circumstances.It would seem, however, that SAS PROC MIXED, although viable, may be inefficient due to computational demands.

Applications

There many situations in educational and behavioral research in which multiple dependent variables are of interest. Oftentimes these variables may take the pattern of path analytic model, but there are many other cases where they do not. However, it is commonplace for educational researchers to conduct separate analyses for multiple dependent variables even though they are likely to be correlated and have similar although not identical design matrices. For example, researchers in counseling often have multiple outcomes (measure of symptoms, coping, etc.) that are assumed to have some of the same predictors but

to also have predictors that are unique to each measure. This is a situation that calls for a SUR model; however, a search of ERIC and PSYCHINFO located 11 applications of SUR models despite the enormous number of articles that analyze multiple dependent variables (see Appendix A). SUR models are underutilized and should be give more consideration as an analytic technique. The issue begins with education, and thus, we as statistics educators should devote more time to covering SUR models as a flexible analytic method in our multivariate analyses courses.

References

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963-74.

Reinsel, G. (1984). Estimation and prediction in a multivariate random effects generalized linear model. *J. Amer. Statist. Assoc.*, *79*, 406-414.

Park, T. (1993). Equivalence of maximum likelihood estimation and iterative two-sage estimation for seemingly unrelated regression. *Communications in Statistics: Theory & Methods*, *22*, 2285-2296.

Schafer J. L, & Yucel, R. M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values. *Journal of Computational and Graphical Statistics*, *11(2)*, 437-457.

Shah, A., Laird, N., Schoenfeld, D. (1997). A Random-Effects Model for Multiple Characteristics With Possibly Missing Data, *Journal of the American Statistical Association*, *92*, 775-779.

Timm, N. H. (2002). *Applied multivariate analysis*. New York: Springer.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association*, *57*, 348-368.

Appendix A
References for Behavioral Science Articles using SUR

Godwin, D. D. (1985). Simultaneous Equations Methods in Family Research. *Journal of Marriage & the Family*, 47, 9-23.

Grofman, B., &; Migalski. (1988). Estimating the Extent of Racially Polarized Voting in Multicandidate Contests. *Sociological Methods & Research*, 16, 427-454.

Gruenewald, P. J. (1997). Analysis approaches to community evaluation. *Evaluation Review*, 21, 209-30.

Fernandez, S., Smith, C. R., & Wenger, J. B. (2007). Employment, Privatization, and Managerial Choice: Does Contracting out Reduce Public Sector Employment? *Journal of Policy Analysis & Management*, *26*, 57-77.

Hook, J. L. (2004). Reconsidering the Division of Household Labor: Incorporating Volunteer Work and Informal Support. *Journal of Marriage & the Family*, *66*, 101-118.

Lehrer, E. L. (1986). Simultaneous Equations Methods in Family Research: A Comment. *Journal of Marriage & the Family*, *48*, 881-883.

Minnotte, K. L., Stevens, D. P., Minnotte, M. C., & Kiger, G. (2007). Emotion-Work Performance among Dual-Earner Couples: Testing Four Theoretical Perspectives. *Journal of Family Issues*, *28*(*6*), 773-793.

Pelled,L. H., Eisenhardt, K. M., & Xin, K. R. (1999). Exploring the black box: An analysis of work group diversity, conflict, and performance. *Administrative Science Quarterly*, *44*(*1*), 1-28.

Price-Spratlen, T. (1998). Between depression and prosperity? Changes in the community context of historical African American migration. *Social Forces*, *77*, 515-539.

Schwartz, J. (2006). Effects of Diverse Forms of Family Structure on Female and Male Homicide. *Journal of Marriage & the Family*, *68*, 1291-1312.

Wu, Z. (2005). Generalized Linear Models in Family Studies. *Journal of Marriage & Family*, *67*, 1029-1047.

Zax, J. S. (2002). Comment on "Estimating the Extent of Racially Polarized Voting in Multicandidate Contests" by Bernard Grofman and Michael Migalski. *Sociological Methods & Research*, *31*, 75-87.

Send correspondence to: T. Mark Beasley
University of Alabama at Birmingham
Email: mbeasley@uab.edu

Appendix B
(SAS/IML Code to Generate Data for Figure 1)

```
proc iml;
N=5000;** sample Size **;
r_e12=0.10;r_e13=0.20;r_e23=0.10;** Correlation among Error terms **;
Rey=(1||r_e12||r_e13)//(r_e12||1||r_e23)//(r_e13||r_e23||1);
r_x12=0.30;r_x13=0.25;r_x23=0.15;** Correlation among Exogeneous Variables **;
Rxx=(1||r_x12||r_x13)//(r_x12||1||r_x23)//(r_x13||r_x23||1);
b1_y2=0.25;b1_y3=.35;********** Path Coefficients for Y1 **;
b1_x2=0.20;******************** Path Coefficients for Y1 **;
b2_x1=0.35;b2_x2=0.20;b2_x3=0;* Path Coefficients for Y2 **;
b3_x1=0;b3_x2=0.40;b3_x3=0.25;* Path Coefficients for Y3 **;
R2_y3=(b3_x1||b3_x2||b3_x3)*Rxx*((b3_x1||b3_x2||b3_x3)`);
R2_y2=(b2_x1||b2_x2||b2_x3)*Rxx*((b2_x1||b2_x2||b2_x3)`);
Rxxe=(Rxx||(j((nrow(Rxx)),1,0)))//(((j(1,(nrow(Rxx)),0)))||1);
vecre23=(0||0||0||r_e23);Rxxe=Rxxe//vecre23;Rxxe=Rxxe||((vecre23`)//1);
R_y23=(b2_x1||b2_x2||b2_x3||((1-R2_y2)##.5)||0)
*Rxxe*((b3_x1||b3_x2||b3_x3||0||((1-R2_y3)##.5))`);
R_y2x2=(b2_x1||b2_x2||b2_x3)*Rxx[,2];
R_y3x2=(b3_x1||b3_x2||b3_x3)*Rxx[,2];
Rxy1=(1||R_y23||R_y2x2)//(R_y23||1||R_y3x2)//(R_y2x2||R_y3x2||1);
print 'Rxx Correlation matrix for Y1' ;print Rxy1;
R2_y1=(b1_y2||b1_y3||b1_x2)*Rxy1*((b1_y2||b1_y3||b1_x2)`);
Rxy1e1=1||(((1-R2_y2)##.5)#r_e12)||(((1-R2_y3)##.5)#r_e13);
Rxy1e1=Rxy1e1//( (((1-R2_y2)##.5)#r_e12)||1||0);
Rxy1e1=Rxy1e1//( (((1-R2_y3)##.5)#r_e13)||0||1);
print 'Correlation Matirx for Y2-Y3-X2';print Rxy1e1;
R_y1e1=(b1_y2||b1_y3||b1_x2)*Rxy1*((b1_y2||b1_y3||b1_x2)`);
ry1e1=((((1-R2_y2)##.5)#r_e12)||(((1-R2_y3)##.5)#r_e13)||0)
*((b1_y2||b1_y3||b1_x2)`);ry1e1=ry1e1/(R2_y1##.5);
print 'Correlation of Y1-e1' ry1e1;
print 'R-squares';print R2_y1 R2_y2 R2_y3 R_y23 R_y2x2 R_y3x2;
seed=13;** Setting Seed gives the same Result everytime ;
*** For Errors of Y ******************;
lame=eigval(rey);*** LATENT ROOTS OF rey  ******************;
lsqrte=diag(lame##0.5);** DIAGONAL MATRIX WITH THE SQUARE ROOT OF EIGENVALUES;
eve=eigvec(rey);** EIGENVECTORS OF rey  ************************;
fre=eve*lsqrte;** CREATE FACTOR SCORE MATRIX (fre)      ******;
Ze= rannor(j(N,3,seed));Ze=fre*Ze`;Ze=Ze`;
*** For X variables ******************;
lamx=eigval(rxx);***  LATENT ROOTS OF Rxx  ******************;
lsqrtx=diag(lamx##0.5);** DIAGONAL MATRIX WITH THE SQUARE ROOT OF EIGENVALUES;
evx=eigvec(rxx);** EIGENVECTORS OF Rxx  ************************;
frx=evx*lsqrtx;** CREATE FACTOR SCORE MATRIX (frx)   ******;
Zx= rannor(j(N,3,0));Zx=frx*Zx`;Zx=Zx`;
*********************************;
e1=Ze[,1];e2=Ze[,2];e3=Ze[,3];
x1=Zx[,1];x2=Zx[,2];x3=Zx[,3];
y3=(b3_x1#x1)+(b3_x2#x2)+(b3_x3#x3)+(((1-R2_y3)##.5)#e3);
y2=(b2_x1#x1)+(b2_x2#x2)+(b2_x3#x3)+(((1-R2_y2)##.5)#e2);
qb=-2#(R2_y1##.5)#ry1e1;** Define the qb coefficient for Quadratic Equation *;
m=(qb+(((qb##2)-(4#(R2_y1-1)))##.5))/2;* Solve positive root of Quad Eq. ****;
print 'Coefficient for e1' m;
y1=(b1_y2#y2)+(b1_y3#y3)+(b1_x2#x2)+(m#e1);
dats=y1||y1a||y2||y3||x1||x2||x3||e1||e2||e3;
varname={'y1'  'y2' 'y3' 'x1' 'x2' 'x3' 'e1' 'e2' 'e3'};
create outs from dats [colname=varname];
   append from dats;
```

Appendix C
(SAS Code to Perform SUR Model and Path Analyses of Data from Figure 1)

```
data outs;set outs;id=_n_;run;
proc corr data=outs;run;
proc standard data = outs out=surpath mean=0 std=1;var y1 y2 y3 x1 x2 x3;run;
proc syslin data=surpath FIML; ** OTHER OPTIONS include SUR and ITSUR ***;
endogenous  y1 y2 y3;          **             INSTEAD of FIML         ***;
instruments x1 x2 x3;
      y1: model y1 = y2 y3    x2    / noint stb;
      y2: model y2 =       x1 x2    / noint stb;
      y3: model y3 =          x2 x3 / noint stb;run;
proc calis data=surpath method=ML;
LINEQS
      y3 = b3_x2 X2 + b3_x3 X3 + e_3,
      y2 = b2_x1 X1 + b2_x2 X2 + e_2,
      y1 = b1_y2 Y2 + b1_y3 Y3 + b1_x2 X2 + e_1;
STD X1=v_x1, X2=v_x2, X3=v_x3, e_3=v_e3, e_2=v_e2, e_1=v_e1;
COV e_1 e_2 = c_e12, e_1 e_3 = c_e13, e_2 e_3 = c_e23; run;
data stack;set surpath; ** STACKING THE DATA for PROC MIXED   ******;
do mod = 1 to 3;
  if mod = 1 then do;
   y=y1;b1_0=1;b1_y2=y2;b1_y3=y3;b1_x2=x2;
        b2_0=0;b2_x1= 0;b2_x2= 0;
        b3_0=0;b3_x2= 0;b3_x3= 0;
   output;
  end;
  if mod = 2 then do;
   y=y2;b1_0=0;b1_y2= 0;b1_y3= 0;b1_x2= 0;
        b2_0=1;b2_x1=x1;b2_x2=x2;
        b3_0=0;b3_x2= 0;b3_x3= 0;
   output;
  end;
  if mod = 3 then do;
   y=y3;b1_0=0;b1_y2= 0;b1_y3= 0;b1_x2= 0;
        b2_0=0;b2_x1= 0;b2_x2= 0;
        b3_0=1;b3_x2=x2;b3_x3=x3;
   output;
  end;
end;
run;
proc mixed data=stack method=ML   ;class mod id;
    model y =  b1_y2 b1_y3 b1_x2
               b2_x1 b2_x2
               b3_x2 b3_x3  /noint solution DDFM=KENWARDROGER   ;
    repeated mod / type=un subject=id r rcorr;run;
```

Mallow's C_p for Selecting Best Performing Logistic Regression Subsets

Mary G. Lieberman **John D. Morris**
Florida Atlantic University

Mallow's C_p is used herein to select maximally accurate subsets of predictor variables in a logistic regression. Across a wide variety of data sets, an examination of the cross-validated prediction accuracy, posited as the ultimate criterion for model performance, contrasts the leave-one-out performance of Mallow's C_p selections with the accuracy afforded by optimal subsets. Losses in accuracies ranged from no loss in several data sets up to a maximum of 10%. The performance of C_p selected subsets can be viewed as promising. It is posited that one should also consider parsimony and the richness of multiple optimal models.

This study investigates the proposition by Hosmer and Lemeshow (2000) that Mallow's C_p be used to select subsets of maximally accurate predictor variables in a logistic regression. As accurate cross-validated prediction accuracy is considered the ultimate criterion for prediction model performance, an examination, across a wide variety of data sets, of the leave-one-out performance of Mallow's C_p selected subsets (in respect to the accuracy of the optimal subset) is examined.

Multiple regression is so thoroughly entrenched in statistical methods that it hardly needs an introduction herein, and is, thus, an obvious modeling technique used to examine the predictive accuracy of subsets of variables. Among the techniques used for solving classification problems, logistic regression (LR) and predictive discriminant analysis (PDA) are two of the most popular (Yarnold, Hart & Soltysik, 1994). Unlike PDA, LR captures the probabilistic distribution embedded in a categorical outcome variable, avoids violations to the assumption of homogeneity of variance, and does not require strict multivariate normality. Therefore, when PDA assumptions are violated, we might expect greater cross-validated classification accuracy with LR than PDA.

Although several studies have compared the classification accuracy of LR and PDA, the results have been inconsistent. For example, some studies (Baron, 1991; Bayne, Beauchamp, Kane, & McCabe, 1983; Crawley, 1979) suggest that LR is more accurate than PDA for nonnormal data. However, several researchers (e.g., Cleary & Angel, 1984; Knoke, 1982; Krzanowski, 1975; Lieberman & Morris, 2003; Meshbane & Morris, 1996; Press & Wilson, 1978) found little or no difference in the accuracy of the two techniques with PDA often performing better than LR. Part of the reason these results are in dispute is that one may consider accuracy for all groups or separate-groups. As well, one may consider a cross-validated index of accuracy or the accuracy of reclassifying the calibration sample; these studies are not consistent in respect to the criterion of accuracy used. Specifically, examination of cross-validation accuracy in LR studies is uncommon, and when done is usually of the most basic (and unstable) sort (hold-out sample). No commercial computer packages support more appropriate resampling cross-validation methods (variously called PRESS, Lachenbruch U, leave-one-out, jackknife and the bootstrap).

Whichever method (LR or PDA) is selected, one may consider subsets of all possible variables for purposes or parsimony, or to *increase* cross-validation accuracy of the model (Morris & Meshbane, 1995). The most usual method is to consider accuracy in classification of the sample upon which the model is created (internal) with the objective of parsimony. That is, realizing that some accuracy will be lost in reducing the number of predictor variables in classifying the calibration sample, but compromising that loss with the gain in parsimony by the reduction in size of the prediction model. However, as in multiple regression, an increase in cross-validated prediction accuracy (the most appropriate criterion) is almost always available using a model composed of fewer than all available variables. Thus one may gain both parsimony and some degree of explanatory power for the model. In addition, although traditional methods considering the piecemeal change in performance of models in respect to prediction within the calibration sample have often been used (forward, backward, stepwise, or variants thereof), they are neither optimal, nor unique, and are now generally in disfavor.

In the case of PDA an examination of the cross-validation accuracy of all 2^p-1 (where p is the number of predictor variables) subsets of variables has been recommended and utilized (Huberty, 1994; Huberty & Olejnik, 2006; Morris & Meshbane, 1995). In this case the method of cross-validation is the leave-one-out method. In the leave-one-out procedure (Huberty, 1994, p. 88; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968) a subject is classified by applying the rule derived from all subjects except the

one being classified. This process is repeated round-robin for each subject, with a count of the overall classification accuracy used to estimate the cross-validated accuracy. Clearly the same round-robin procedure can be used to estimate either relative or absolute accuracy in the use of multiple regression and has appeared in that context, with perhaps the earliest reference due to Gollob (1967). In a system intended to select optimal multiple regression predictor variable subsets, Allen (1971) coined the procedure *PRESS*, and he appears to be the source most often cited in the multiple regression literature.

In the case of PDA (and regression) a matrix identity due to Bartlett (1951) allows the task of N-1 discriminant analyses to be accomplished with far less computational labor that would otherwise be necessary. However, this mathematical tool is irrelevant to the iterative method of LR optimization, thus N-1 LR optimizations must be completed for each of 2^p-1 subsets of predictor variables.

Unlike most LR studies that consider calibration sample statistics as the criterion for model fit (e.g., the Cox & Snell, or Nagelkerke R^2), the criterion for model accuracy is construed in this study, as is typically done in PDA, as classification accuracy - that is, the proportion of correct leave-one-out cross-validated classifications (hit-rate) for the total sample and each separate group. Thus for a two-group problem, we order the accuracy of our 2^p-1 candidate LR equations according to three different (total sample and each group) cross-validated classification accuracy criteria.

An alternative logistic regression variable selection strategy has been proposed by Hosmer and Lemeshow (2000) using a technique due to C. L. Mallows (1973). Although Mallows' technique was intended for OLS regression variable subset selection, with attendant consideration of its merit in that context (e.g., Schumacker, 1994), the direct suggestion of Hosmer and Lemeshow of its use in variable subset selection in logistic regression is directly examined herein.

Methods

Analyses from 19 two-group classification problems from Morris and Huberty (1987) were used in this comparison. Although not purported to represent all potential data structures, these data sets have been used in several classification studies as representing a wide variety of number of predictor variables, group separation, and covariance structures.

For a variety of data sets the leave-one-out cross-validated classification accuracies for the Mallows C_p selected variable subset was compared to that derived from the subset manifesting maximum classification accuracy. The difference between the maximum hit-rate and number of predictors for the best subset and that selected by Mallows C_p was compared. The criterion of model accuracy in this study is the proportion of correct leave-one-out cross validated classifications (hit rate) for the total sample and each separate group.

Results & Discussion

Table 1 shows the data source, number of predictors for the full model, hit-rate for the full model, number of predictors in the best subset (s), and maximum hit-rate in the first five columns from left to right. For Mallows C_p, the final three columns show the number of predictors in the C_p selected subset, the hit-rate for that subset, and the percentage loss in hit-rate from the best subset chosen from the maximum hit-rate.

In all cases selection of the best performing subset (of the 2^p-1 possibilities) offers a reduction in the number of predictor variables, often by more than half, thus parsimony is well served. In the first five data sets there is no loss in hit-rate accuracy and equal parsimony using Mallows C_p as with respect to all possible subsets. In data sets numbered seven and fifteen there is no loss in hit-rate accuracy, although the most parsimonious subset is not selected by C_p. In the remaining data sets, losses in accuracy incurred by use of the C_p strategy ranged from .97% – 10.60%.

In several cases one can have enhanced parsimony, hit-rate accuracy close to maximum, and reduced computational intensity using Mallows C_p as the predictor variable selection procedure. The performance of Mallows C_p could be viewed as promising.

Another use of the consideration of the accuracy of all possible subsets involves the treatment of missing data. Table 2 demonstrates the potential use of several alternative "best" models. These data represent the top twenty best subsets of variables in an 8th grade dropout profiling study including

Table 1. Data set, # variables (*p*), Hit rate for all, Maximum, and C_p selected and % Loss.

#	Data Set Source	*p*	Hit-rate for p Predictors	# Predictors in Best Subset(s)	Maximum Hit-rate	C_p # Predictors	C_p Hit-Rate	% Loss
1	Rulon Grps 1 & 2	4	0.803	3	.815	3	.815	**0.00**[a]
2	Rulon Grps 1 & 3	4	0.914	3	.934	3	.934	**0.00**
3	Rulon Gps 2 & 3	4	0.824	3	.830	3	.830	**0.00**
4	Block - Grps 1 & 2	4	0.692	1,2	.718	1	.718	**0.00**
5	Block - Grps 1 & 3	4	0.620	3,4	.620	3	.620	**0.00**
6	Block - Grps 1 & 4	4	0.577	1,2	.628	2	.615	0.02
7	Block - Grps 2 & 3	4	0.566	1,2	.605	2	.605	0.00
8	Block - Grps 2 & 4	4	0.587	2	.627	1	.587	6.37
9	Block - Grps 3 & 4	4	0.684	3	.697	1	.632	9.32
10	Demographics	8	0.591	4	.620	3	.609	1.77
11	Dropout from 4th	10	0.660	4	.787	4	.681	10.60
12	Dropout from 8th	11	0.725	3	.782	4	.746	4.60
13	Fitness	10	0.591	4	.620	4	.588	5.16
14	Warncke-Grps 1 & 3	10	0.600	4	.667	3	.619	7.19
15	Bisbey 1& 2	13	0.879	6,7,8,9,10	.914	9	.914	0.00
16	Bisbey 2& 3	13	0.856	5,6,7	.924	3	.915	.97
17	Talent - Grps 1 & 3	14	0.621	5	.733	2	.707	3.54
18	Talent - Grps 3 & 5	14	0.787	6,7,8,9	.858	7	.811	5.47
19	Talent - Grps 1 & 5	14	0.740	5	.797	7	.751	5.77

[a] Bold denotes equal performance and parsimony.

Table 2. Ranked 20 best (of 255) performing subsets, and total model.

HIT-RATE	SCHOOLS8	REPEATS8	READING8	MATH8	LANG8	SCIENCE8	SOCST8	DSFS8
0.753	√			√		√		√
0.747	√			√				√
0.747	√							√
0.747	√				√	√	√	√
0.747	√			√	√	√	√	√
0.741	√		√		√		√	√
0.741	√			√		√	√	√
0.735	√		√		√	√		√
0.735	√			√			√	√
0.735	√	√	√				√	
0.735	√			√	√		√	√
0.735	√					√	√	√
0.735	√		√	√	√			√
0.735	√	√						√
0.735	√	√	√					
0.728	√	√				√	√	
0.728	√					√		√
0.728	√	√				√		√
0.728	√	√	√			√		√
0.728	√	√				√		
Total Model								
0.679	√	√	√	√	√	√	√	√

number of schools attended by the 8th grade, standardized test scores, and the number of D's and F's obtained during the 8th grade year. Depending on which variables are missing for a subject, with knowledge of the best performing subsets, it may be possible to select a superior subset appropriate for data that a subject has available. An advantage to looking at all possible subsets is the allowance for the elimination of variables for which numbers of subjects are missing data.

The table shows a check mark if a variable appears in each of the top twenty models (out of two hundred and fifty five). Considering column-wise entries, a frequent notion of variable importance seems appropriate. When parsimony and accuracy are considerations for model fit, it is clear from these data that, for example, schools attended by 8th grade is a 'don't leave home without it' variable, as it appears in all of the top twenty models. Similarly, Number of D's and F's obtained by eighth grade appears in most models as does number of science courses taken by eighth grade. The other variables, although desirable, may demonstrate little adequacy, in an additive sense, for inclusion in a prediction model. Therefore, this view of variable importance is such that since some variables appear in all or most models, one might suggest this as a defensible measure of variable importance.

In this particular case, since current emphases on standardized testing, and other indices of achievement, tend to focus on predicting success and profiling students at risk, while lessening the drain on time consumption and fiscal resources, such a measure of variable importance may be considered a vital aspect of any prediction formula.

References

Allen, D. A. (1971). *The prediction sum of squares as a criterion for selecting predictor variables* (Tech. Rep. No. 23). Lexington: University of Kentucky, Department of Statistics.

Baron, A. E. (1991). Misclassification among methods used for multiple group discrimination – The effects of distributional properties. *Statistics in Medicine*, *10*, 757-766.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *Annals of Mathematical Statistics*, *22*, 107-111.

Bayne, C. K., Beauchamp, J. J., Kane, V. E., and McCabe, G. P. (1983). Assessment of Fisher and logistic linear and quadratic discrimination models. *Computational Statistics and Data Analysis*, *1*, 257-273.

Cleary, P. D. & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior*, *25*, 334-348.

Crawley, D. R. (1979). Logistic discriminant analysis as an alternative to Fisher's linear discriminant function. *New Zealand Statistics*, *14(2)*, 21-25.

Gollub, H. F. (1967, September). *Cross-validation using samples of size one*. Paper presented at the annual meeting of the American Psychological Association, Washington, D. C.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

Huberty, C. J, & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. New York: Wiley.

Huberty, C. J (1994). *Applied discriminant analysis*. New York: Wiley.

Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, *38*, 191-200.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, *70*, 782-790.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10,* 1-11.

Lieberman, M. G, & Morris, J. D. (2003, April). *Comparing classification accuracies between predictive discriminant analysis and logistic regression in specific data sets*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*, 661-675.

Meshbane, A., & Morris, J. D. (1996, April). *Predictive discriminant analysis versus logistic regression in two-group classification problems*. Paper presented at he meeting of the American Educational Research Association, New York.

Morris, J. D., & Huberty, C. J (1987). Selecting a two-group classification weighting algorithm. *Multivariate Behavioral Research, 22*, 211-232.

Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. *Educational and Psychological Measurement, 55*, 438-441.

Mosteller, F. & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.). *Handbook of social psychology* (Vol. 2, pp. 80-203). Reading, MA: Addison-Wesley.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*, 699-705.

Schumacker, R. E. (1994). A comparison of the Mallows C_p and principal component regression criteria for best model selection in multiple regression. *Multiple Linear Regression Viewpoints*, 21, 12-22.

Yarnold, P. R., Hart, L. A. & Soltysik, R. C. (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analysis. *Educational and Psychological Measurement*, *54*, 73-78.

Send correspondence to: Mary G. Lieberman
Florida Atlantic University
Email: mlieberm@fau.edu

Regression Discontinuity Models and the Variance Inflation Factor

Randall E. Schumacker
University of Alabama

The regression-discontinuity design (RD) is a powerful methodological alternative to the quasi-experimental design when conducting evaluations. The RD design involves testing post-test differences between the experimental and comparison group regression lines at the cutoff point for statistical significance. Regression discontinuity models can involve linear, curvilinear, and interaction terms in the model specification, which are not orthogonally specified. Consequently, a variance inflation problem may exist when using regression discontinuity models in evaluation designs. This study investigated the impact of variance inflation on parameters specified in full and restricted regression discontinuity models. It is recommended that VIF be considered when including interaction effects in RD designs.

The basic RD Design is a two-group pretest-posttest model and is depicted as follows:

C O X O
C O O

The RD design looks similar to the Non-Equivalent Group design, which uses analysis of covariance, but assumptions and advantages are much different. The RD design does not have subject selection bias (pre-defined group membership) rather uses a pre-test measure to assign treatment or non-treatment status. The basic RD model would have an intercept term, pre-test measure, and dummy-coded group assignment variable regressed on a post-test measure. The pre-test measure does not have to be the same as the post-test measure.

There are five central assumptions when performing an RD analysis. These are:

1. The cutoff value must be absolute without exception. A subject selection bias is introduced and the treatment effect is biased if incorrect assignment to groups based on the cutoff value occurred (unless it is known to be random).
2. The pre-post distribution is a polynomial function. If the pre-post relationship is logarithmic, exponential or some other function, the model is misspecified and the treatment effect is biased. The data can be transformed to create a polynomial distribution prior to analysis to yield appropriate model specification.
3. There must be a sufficient number of pretest values in the comparison group to estimate the pre-post regression line.
4. The experimental and comparison groups must be formed from a single continuous pretest distribution with the division between groups determined by the cutoff value.
5. The treatment or program intervention must be delivered to all subjects, i.e., all receive the same reading program, amount of training, etc.

Regression Discontinuity Model Specification

The major concern when analyzing data from the RD design is whether the model or regression equation is correctly specified. If the regression equation or model does not reflect the data distribution, then biased estimates of the treatment effect will occur. For example, if the true pre-post relationship is curvilinear, but the regression equation only modeled linear regression effects, the treatment effects would be biased. Consequently, it is a good idea to visually inspect the pre-post scatter plot to see what type of relationship exists.

Three types of model specifications are possible: exactly specified, over specified, and under specified RD models. An exactly specified model has an equation that fits the "true" data. So if the "true" data is linear then a simple straight-line pre-post relationship with a treatment effect would yield unbiased treatment effects. The RD equation would include a term for the posttest Y, the pretest X, and the dummy-coded treatment variable Z with no unnecessary terms. When we exactly specify the true model, we get unbiased and efficient estimates of the treatment effect. If the RD equation is over specified it includes additional parameter estimates that are not required, i.e. interaction or curvilinear coefficients, and treatment effect would be inefficient. If the RD equation is under specified it leaves out important parameter estimates and the treatment effect would be biased.

RD Modeling Steps

The basic steps to conducting RD analyses would as follows:

1. Subtract the cut-off score from the pretest score ($X_{pre} - X_{cut}$).
2. Visually examine the pre-post scatter plot for type of data relationship.
3. Determine if any higher-order polynomial terms or interactions are present.
4. Estimate the "full" RD regression equation.
5. Modify the RD equation by dropping individual non-significant terms.

The "full" RD regression equation with subsequent "modified" or "restricted" regression models permit one to statistically determine the best fitting model for estimating treatment effects. A "full" regression discontinuity model could be as outlined below.

$$y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + \beta_5 X_i^2 Z_i + e_i$$

The RD regression equation terms are defined as:

y_i = post test score outcome for i^{th} subject
β_0 = regression coefficient for intercept
β_1 = linear pre test regression coefficient
β_2 = mean post test different for treatment group
β_3 = linear interaction regression coefficient between pre and group
β_4 = quadratic regression coefficient for pretest
β_5 = quadratic interaction regression coefficient for pre test and group
X_i = transformed pre test score for i^{th} subject
Z_i = group assignment based on cut off score (0 = comparison, 1 = treatment)
e_i = residual score for i^{th} subject.

Variance Inflation Factor

When a full RD regression model is specified, multicollinearity amongst the terms is possible. Multicollinearity can inflate the variance amongst the variables in the model. These inflated variances are problematic in regression because some variables add very little or even no new and independent information to the model (Belsley, Kuh & Welsch, 1980). Although Schroeder, Sjoquist and Stephen (1986) assert that there is no statistical test that can determine whether or not multicollinearity is a problem, there are ways for detecting multicollinearity (Berry and Feldman, 1985).

A recommended approach is to use the Variance Inflation Factor (VIF). VIF measures the impact of multicollinearity among the *X*'s in a regression model on the precision of estimation. It expresses the degree to which multicollinearity amongst the predictors degrades the precision of an estimate. VIF is a statistic used to measured possible multicollinearity amongst the predictor or explanatory variables. VIF is computed as $1/(1-R^2)$ for each of the k–1 independent variable equations. For example, given 4 independent predictor variables, the independent regression equations are formed by using each k-1 independent variable as the dependent variable:

$X_1 = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_4 + e_1$
$X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + e_2$
$X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_3$

Each independent variable model will return an R^2 value and VIF value. The term to exclude in the model is then based on the value of VIF. If X_j is highly correlated with the remaining predictors, its variance inflation factor will be very large. A general rule is that the VIF should not exceed 10 (Belsley, Kuh, & Welsch, 1980). When X_j is orthogonal to the remaining predictors, its variance inflation factor will be 1.

Methods

Data Simulation

The appendix contains an S-PLUS program that generated the simulated data for the study. The rnorm function in S-PLUS generated 100 random normal data points and output nine variables listed in the data command [data <-c(y,x,z,gain,ypost,xc,xz,xsq,xsqz)]. The post test scores (*Y*) and pre test scores (*X*) were created by adding residual error (*ey* or *ex*) to this random normal variable (*true*). Group assignment (*Z*) was determined based on subtracting a cut score of 20 from the pre test score (1–treatment, 0–comparison). This 10 point treatment gain was added to the post test score (*Y*). Optional *print* and *write* statements are included to either view or save the data in a file.

Regression Discontinuity Models

The least squares regression function, *lm*, was used to run the RD analyses. The S-Plus program includes separate *lm* regression functions for several regression equations. The summary command produced the regression output. The regression discontinuity models begin with a full model followed by a sequence of restricted models. The full regression model and the sequence of restricted models are listed below:

1. Full model: $y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + \beta_5 X_i^2 Z_i + e_i$
2. No Quadratic Interaction: $y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + \beta_4 X_i^2 + e_i$
3. No Quadratic Terms: $y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + \beta_3 X_i Z_i + e_i$
4. Linear Model: $y_i = \beta_0 + \beta_1 Z_i + \beta_2 X_i + e_i$
5. No Pre-Test Model: $y_i = \beta_0 + \beta_1 Z_i + e_i$

Variance Inflation Factor

The variance inflation factor is computed in several popular statistics packages (S-PLUS, SPSS, and SAS). In this study, the data simulation, regression function, and variance inflation function were all written in S-PLUS. The simulated data generated using the S-PLUS program in the appendix was created and used by the S-PLUS regression and variance inflation functions. A variance inflation function, *vif,* was created and used with the *summary* function following the *lm* regression function for each of the regression equations. The S-PLUS variables were labeled as follows in the full regression equation: ypost = xc + z + xz + xsq + xsqz.

Results

The descriptive statistics for the RD variables are in Table 1. The intercorrelations amongst the terms in the full RD regression model equation are in Table 2. The RD regression discontinuity results with the VIF values for the full model are in Table 3 for the dependent variable ypost.

Table 1. Descriptive Statistics (*N*=100)

	Mean	Std. Deviation
ypost	25.5852	5.45566
xc	.0899	1.35059
z	.55	.500
xz	11.5740	10.53759
xsq	405.4103	53.68331
xsqz	243.8876	223.08339

Table 2. Pearson Correlation Matrix of Full RD Regression Model

	ypost	xc	z	xz	Xsq	xsqz
ypost	1.000	.821	.971	.971	.821	.969
xc	.821	1.000	.785	.807	.999	.827
z	.971	.785	1.000	.999	.787	.994
xz	.971	.807	.999	1.000	.811	.998
xsq	.821	.999	.787	.811	1.000	.833
xsqz	.969	.827	.994	.998	.833	1.000

Table 3. Full Regression Model and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	-9.772	57.928	-.168	.866		
	xc	-1.909	5.298	-.360	.719	.000	3467.21
	z	-87.861	108.980	-.806	.422	.001	201043.10
	xz	9.978	10.576	.943	.347		841002.30
	xsq	.076	.145	.526	.600	.000	4131.94
	xsqz	-.257	.258	-.993	.323	.001	225640.50

Table 4. Restricted Regression Model (no xsqz) and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	22.580	47.898	.471	.638		
	xc	1.048	4.382	.239	.812	.000	2372.36
	z	19.149	16.345	1.172	.244	.000	4523.18
	xz	-.493	.821	-.601	.549	.000	5067.20
	xsq	-.005	.120	-.039	.969	.000	2825.33

Table 5. Restricted Regression Model (no xsq) and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	20.703	.279	74.128	.000		
	xc	.876	.199	4.401	.000	.202	4.95
	z	19.748	5.807	3.401	.001	.002	576.84
	xz	-.523	.289	-1.810	.073	.002	635.16

Table 6. Restricted Regression Model (no xz) and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	20.436	.240	85.160	.000		
	xc	.628	.146	4.300	.000	.384	2.60
	z	9.259	.395	23.471	.000	.384	2.60

Table 7. Restricted Regression Model (no z) and VIF

Model		Unstandardized Coefficients		t	Sig.	Collinearity Statistics	
		B	Std. Error			Tolerance	VIF
1	(Constant)	25.287	.314	80.643	.000		
	xc	3.317	.233	14.249	.000	1.000	1.00

Summary & Conclusion

The requirement of a correctly specified RD regression model is linked to multicollinearity of the independent variables in the equation. Table 2 suggests that multicollinearity is present amongst the independent predictors in the RD regression equation, i.e. β_1 = linear pre test regression coefficient (xc); β_2 = mean post test different for treatment group (z); $\beta 3$ = linear interaction regression coefficient between pre and group (xz); β_4 = quadratic regression coefficient for pretest (xsq); and β_5 = quadratic interaction regression coefficient for pre test and group (xsqz). Table 3 indicates that VIF is well beyond the acceptable level of 10 for each of the independent predictor variables in the model. Similar results were

found for the set of independent predictor variables in Table 4, especially note the non-significant treatment effect (z) with an extreme VIF factor. Table 5 indicated that the linear pre test regression coefficient (xc) was acceptable, however, the other independent predictors VIF were too high, i.e., the treatment effect is now significant, but has an extreme VIF factor. In Table 6, a two predictor model with linear pre test and treatment group had both a significant *t*-test value (t = 23.471, p = .0001) and an acceptable VIF factor; thus an acceptable RD model. Table 7, indicated a baseline RD model with linear pre test scores and an expected corresponding VIF = 1.0.

The regression discontinuity approach to analyzing evaluation data is more robust to violations than the corresponding quasi-experimental design that is commonly used in state and federal grant data analysis. However, model misspecification can result in erroneous conclusions regarding program gains. Correspondingly, if the variance inflation factor is not considered along with model specification, then multicollinearity amongst the predictor variables can inflate the variance leading to misinterpretation of the R-squared values and treatment gain. A visual presentation of overlap by the independent variables is also possible (Stine, 1995). It is therefore recommended that model specification along with the variance inflation factor be checked when using regression discontinuity.

References

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression Diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley.

Berry, W. D. & Feldman, S. (1985). *Multiple regression in practice*. London: Sage Publications.

Schroeder, L. D., Sjoquist, D. L. & Stephan, P. E. (1986). *Understanding regression analysis*. Beverly Hills, CA: Sage Publications.

Stine, Robert, A. (1995, February). Graphical Interpretation of Variance Inflation Factors. *The American Statistician, 49*, 53-56.

Send correspondence to: Randall E. Schumacker
University of Alabama
Email: rschumacker@ua.edu

APPENDIX
S-PLUS Program

```
#
# Data for Normal Distribution
# Pretest X cutoff score is 20 (mean X)
# Program Gain is 10
# Mean Posttest Y is 30
# XC is Pre test minus cut score to center at 0 point
# ex and ey add residual error to true score
#

seed <-1357
set.seed(seed)              # same seed value so results can be reproduced

true <- rnorm(100,20,1)
ex   <- rnorm(100,0,1)
ey   <- rnorm(100,0,1)

x <- true + ex              # create y and x scores with residual error
y <- true + ey

z <- ifelse(x >= 20, 1, 0)  # assign treatment group using pretest cutoff score

gain <- (10 * z)            # add 10 point to treatment group (z = 1)

ypost <- y + gain           # add 10 points to post test score

xc<- (x-20)                 # subtract cut score from pre test

xz<-x*z                     # linear interaction pre test and group

xsq<-x*x                    # quadratic interaction

xsqz<-xsq*z                 # quadratic interaction pre test and group
```

```
data<-c(y,x,z,gain,ypost,xc,xz,xsq,xsqz)

RD.data<-matrix(data,nrow=100,byrow=F)
dimnames(RD.data)
dim(RD.data)  #100 rows 9 columns
variables<-c("y","x","z","gain","ypost","xc","xz","xsq","xsqz")
dimnames(RD.data)<-list(NULL,variables)

#print(RD.data)
#save generated data in ASCII file
write.table(RD.data, file = "RD.txt", sep=",", append=F)

#
#Variance Inflation Factor Function
#

vif <- function(object, ...)
UseMethod("vif")

vif.default <- function(object, ...)
stop("No default method for vif.  Sorry.")

vif.lm <- function(object, ...)

{
  V <- summary(object)$cov.unscaled
  Vi <- crossprod(model.matrix(object))
        nam <- names(coef(object))
  if(k <- match("(Intercept)", nam, nomatch = F)) {
                v1 <- diag(V)[-k]
                v2 <- (diag(Vi)[-k] - Vi[k, -k]^2/Vi[k,k])
                nam <- nam[-k]
        } else {
                v1 <- diag(V)
                v2 <- diag(Vi)
                warning("No intercept term detected.  Results may surprise.")
        }
        structure(v1*v2, names = nam)
}

#
#RD Regression models with Variance Inflation Factor
#
#Sequence of RD equations
#

fit <- lm (ypost~xc + z + xz + xsq + xsqz)
summary(fit)
vif(fit)

fit <- lm (ypost~xc + z + xz + xsq)
summary(fit)
vif(fit)

fit <- lm (ypost~xc + z + xz)
summary(fit)
vif(fit)

fit <- lm (ypost~xc + z)
summary(fit)
vif(fit)

fit <- lm (ypost~xc)
summary(fit)
vif(fit)
```

Evaluation of the Use of Standardized Weights for Interpreting Results from a Descriptive Discriminant Analysis

W. Holmes Finch **Teresa Laking**
Ball State University

When conducting descriptive discriminant analysis, many researchers make use of structure coefficients, the correlation between individual predictor variables and a discriminant function. However, previous research has demonstrated that these statistics may lead to an over-identification of variables important for group separation. An alternative to structure coefficients is the standardized discriminant function weights for the individual variables, which can be used to order variables in importance. Relatively little empirical research has been done examining how well they work in this regard. This study examined the utility of standardized weights for interpreting a discriminant function. Results suggest that the standardized weights may be a useful tool for ordering predictor variables and characterizing significant discriminant functions when the assumptions of normality and homogeneity of covariance matrices are met. When these assumptions are violated, the ability of the standardized weights to correctly order predictor variables was somewhat degraded.

Discriminant Analysis (DA) is a commonly used statistical procedure that allows for both a multivariate description of group differences and the prediction of group membership for individual observations based upon a set of predictor variables. In the first context, typically referred to as Descriptive Discriminant Analysis (DDA), the focus of research is on characterizing differences between two or more groups by identifying which variables among a set of predictors most distinguishes among the groups. In contrast, the goal of Predictive Discriminant Analysis (PDA) is to use the predictors as a set for identifying which of the groups an individual is most likely to belong to. While there may be some interest in assessing the relative contribution of the variables to group separation when conducting a PDA, typically the researcher focuses on the accuracy of group prediction and its potential utility for classifying individuals in the future. It should be noted that while the goals of these two types of DA are different, the underlying mathematical model upon which they are built is the same. This model, which is discussed below, is based on the estimation of linear combinations of the predictors that provide maximal group separation for the sample at hand. The specific focus of the current study is on the utility of the standardized weights in DDA for correctly ordering a set of predictor variables in terms of their relative contribution to group separation in the form of statistically significant linear combinations. The organization of the manuscript is as follows: First is a brief description of DDA and the standardized weights used to create the linear combinations. Following this is a discussion of how these standardized weights can be used for ordering variables in terms of their importance in discriminating between groups. Finally, the details of the current simulation study are discussed, followed by the results and discussion of their implications.

As mentioned above, regardless of whether the application involves description or prediction, DA identifies one or more linear combination of the predictor variables that provide maximum group separation. The number of these linear discriminant functions is equal to the smaller of the number of predictor variables or the number of groups – 1. The relative ability of these functions to distinguish between the groups declines from the first through the second and so on. In addition, it should be noted that although it is mathematically possible to have multiple discriminant functions, in practice not all of them need be statistically significant. In other words, some of these functions may not differentiate the groups in a meaningful way. Thus, the first step in interpretation of a DDA analysis is the examination of test statistics (e.g., Wilks' Lambda) indicating which of the functions are statistically significant. Those that are found to be significant can then be interpreted using tools described below. For a more thorough discussion of the various test statistics available for use in such situations, the reader is encouraged to refer to multivariate texts such as Tabachnick and Fidell (2001).

The actual form of the discriminant function appears in equation (1).

$$D_i = d_{i1}z_1 + d_{i2}z_2 + ... + d_{ij}z_j \quad ; \tag{1}$$

where D_i = the standardized score for an individual on discriminant function i, d_{ij} = the standardized discriminant function coefficient for function i and variable j, and z_j = the value of the standardized predictor variable j.

The discriminant weights, d_{ij}, are determined so as to provide the maximum separation possible on the function value, D_i among the groups in question(Tabachnick & Fidell, 2001). Weights are estimated for each of the *i*discriminant functions separately, and a value of D_i is obtained for each function and each individual in the sample. The means of these D_i are known as group centroids, and their relative proximity can be taken as an indication of the multivariate separation among the groups in question (Huberty & Olejnik, 2006).

The standardized discrimination coefficients take the form:

$$d_{ij} = d_{ij}^{*}\sqrt{s_j^2} \tag{2}$$

where d_{ij}^{*} = the unstandardized discriminant coefficient and s_j^2 = the variance for variable *j*.

In turn the unstandardizeddiscriminant function coefficients are estimated by solving equation (3) below for $\mathbf{d}^{*}$.

$$(\mathbf{E}^{-1}\mathbf{H} - \lambda_1\mathbf{I})\mathbf{d}^{*} = 0 \tag{3}$$

where $\mathbf{E}$ = the Error Sums of Squares and Cross Products matrix, $\mathbf{H}$ = the Hypothesis Sums of Squares and Cross Products Matrix, λ_1 = the maximum eigenvalue for the product of $\mathbf{E}^{-1}\mathbf{H}$, $\mathbf{I}$ = an Identity matrix and $\mathbf{d}^{*}$ = a vector of unstandardized discriminant coefficients.

These fundamental equations for DA rely upon three assumptions regarding the data in the population: 1) The predictor variables are normally distributed; 2) The covariance matrices for the groups are homogeneous; and 3) The residuals for individual subjects are independent of one another (Tabachnick & Fidell, 2001). There has been some research published regarding the impact on PDA of violating these assumptions (Finch & Schneider, 2005; Hess, Olejnik & Huberty, 2001; Meshbane & Morris, 1996: McLachlan, 1992). Taken together, results of these prior studies suggest that the accuracy of PDA in terms of correctly placing individuals in the appropriate group was negatively impacted by violations of the assumptions of normality and homogeneity of covariance matrices. Furthermore, the most negative impact was evident when both assumptions were violated simultaneously (Finch & Schneider, 2005). While these studies focused on the performance of PDA, the fact that the underlying model is essentially the same for DDA makes them relevant to the current work. Therefore, one goal of this study is to ascertain the extent to which violations of the normality and homogeneity of covariance matrices assumptions impact the standardized discriminant weights.

Those discriminant functions that have been identified as statistically significant, can be viewed as effectively differentiating the groups in question. Given such a significant outcome, a researcher would very likely want to gain an understanding as to the nature of such differences; i.e. what do each of the predictor variables contribute to the overall discriminant function that has been shown to differentiate the groups (Rencher, 1995)? There are multiple approaches that have been suggested for use in characterizing the functions based upon the contribution of the individual predictor variables, with the two most common being interpretation of the standardized discriminant coefficients and interpretation of structure coefficients (SC's) (Huberty&Olejnik, 2006).

SC's can be interpreted as the correlation between a discriminant function and the individual predictor variables upon which it is based (Huberty & Olejnik, 2006). Researchers have argued that the SC's are appropriate for interpreting DDA because they provide direct information regarding the relationship between a discriminant function that significantly differentiates among the groups and the individual predictors (Stevens, 2000). Thus, if a variable has a high SC, it can be concluded that it is highly associated with group separation. It should be noted that SC's in DDA are similar in concept to factor loadings, which are used routinely in characterizing the nature of latent factors (Huberty&Olejnik, 2006). Therefore, it may be reasonable to use them for characterizing discriminant functions in much the same fashion.

Because they are similar (though not identical) to factor loadings, some authors have suggested applying arbitrary cut-off values for identifying "important" variables, as is commonly done in exploratory factor analysis. For example, Tabachnick and Fidell (2001) have recommended that SC's larger than 0.32 be considered "important" in terms of understanding the nature of the discriminant function. They selected this value because 0.322 is roughly 0.1, indicating that 10% of the variance in the predictor variable is accounted for by the discriminant function. Pedhazur (1997) recommended using a cut value of 0.3, while other authors (e.g., Huberty & Olejnik, 2006; Stevens, 2000) suggest that

researchers not use a single value, but rather focus on the relative magnitude of the SC's, placing greater emphasis on interpreting those variables with larger values. Dalgleish (1994) introduced a bootstrap confidence interval for use with SC's in DDA. He hoped that this approach would obviate the need for applying arbitrary cut off values by providing information regarding whether, in the population, a given SC differs from 0. If this were the case, Dalgleish argued that a practitioner could then know, with some level of confidence, that a given predictor variable was associated with a significant discriminant function.

Researchers have studied the effectiveness of SC's for interpreting significant discriminant functions in DDA. For example, Dalgleish (1994) found that the bootstrap confidence intervals that he developed for SC's had somewhat conservative Type I error rates, but generally did a better job at maintaining Type I error near the nominal 0.05 level than did arbitrary cut values, including 0.3, 0.4 and 0.5. Finch (2007) conducted a simulation study examining both the Type I error rates (incorrectly identifying a predictor variable as "important" in group separation) and power (correctly identifying a predictor variable as "important" in group separation) of various methods for interpreting SC's, including cut values (0.3, 0.4 and 0.5), relative ordering of importance and the bootstrap confidence interval. Results of this study indicated that in general, the use of SC's led to an over identification of variables associated with group separation. In other words, a researcher using any of these approaches for interpreting SC's could expect to conclude that one or more variables are related to the significant discriminant function when in fact they are not. In addition, the Finch study reported that when the assumptions of normally distributed predictors with equal covariance matrices across groups were violated, the Type I error inflation was particularly severe.

Some researchers have long advocated against using SC's for interpreting significant discirminant functions, and in favor of the standardized weights described above (e.g., Rencher, 1992). The argument in favor of this approach, set forth by Rencher (1995), is that the standardized weight for a particular variable reflects its contribution to the discriminant function in the presence of the other predictors. On the other hand, Rencher argued that the SC relating this variable to the discriminant function demonstrates only the univariate contribution of the individual predictor in question, totally ignoring the presence of the others. For this reason, he asserted that "…these correlations are useless in gauging the importance of a given variable in the context of others because they provide no information about how the variables contribute jointly to separation of the groups. Consequently, they become misleading if used for interpretation of discriminant functions" (Rencher, 1995, p. 317). Instead, he argued on behalf of referring to the discriminant weights when interpreting DDA, because they do account for all of the variables in the model and are therefore more appropriate when one is interested in characterizing significant discriminant function results.

This opinion that standardized weights are more appropriate than SC'sfor use in interpreting discriminant functions is not universally shared. Huberty and Wisenbaker (1992) objected to using the weights because, they stated, simply ordering variables in importance does not communicate anything regarding the different degrees of variable importance, only that one is more important than another. Huberty and Olejnik (2006) go on to argue against the notion of ordering variables in terms of relative importance as a generally useful exercise, and instead focus on characterizing the discriminant function by ascertaining which of the predictor variables were most highly correlated with it, based on the SC's.

Clearly, given the discussion above, the disagreement between methodologists regarding the appropriate approach for interpreting significant discriminant functions has not been resolved to date. In addition to the studies described above that focused on SC's, Huberty (1975) also conducted a simulation study in which he compared the ability to identify predictor variables relevant to group separation of standardized weights and SC's. The outcome variable in this study was the consistency of variable ranking in terms of relative contribution to a significant discriminant function. The data were generated from a normal distribution with equal covariance matrices across 3 and 5 groups for 10 predictor variables. Sample sizes were set at 90, 150, 300 and 450. Huberty concluded that in the 5 groups case, the SC's were slightly more effective at ordering the predictors, while in the 3 groups case the standardized weights performed slightly better in this regard. As he stated, these results are limited to the case where groups are of equal size and the assumptions of equality of covariances and normality are met.

In contrast to the Huberty study, Rencher (1992) described analytically why there may be problems with using the SC's to interpret discriminant functions, and in turn why the standardized weights might be preferable. As noted above, he showed that in the 2-groups case, the SC'sare mathematically proportional

to the univariate t-test comparing the means on the predictor variable between the two groups. Thus, he argued, a researcher making use of the SC's has simply taken what is inherently a multivariate problem and reduced it to a series of univariate ones (Rencher, 1992). Rencher concluded his paper by stating that standardized weights, rather than SC's, are most appropriate for interpreting significant discriminant functions because they allow for a direct ordering of individual predictors in terms of importance while accounting for the presence of all of the other predictors.

Based upon prior research examining the performance of SC's (Finch, 2007, Dalgleish, 1994) there remain some doubts regarding their effectiveness in helping researchers interpret significant discriminant functions. Specifically, regardless of the rule used, SC's appear to over-identify the importance of individual variables in terms of their contribution to group separation. In addition, based upon Rencher's (1992) arguments, these SC values may not be addressing the appropriate multivariate question, namely which variables contribute the most to group separation, in the presence of the other variables in the analysis? Given these potential problems with SC's described by Rencher and highlighted in prior simulation studies, and the relative lack of Monte Carlo research examining the performance of standardized weights in characterizing group differences in DDA, the primary goal of the current study was to use simulations to ascertain how well the standardized weights could order variables in terms of relative importance in group separation under a variety of conditions, which are outlined below. It is hoped that this effort will add to the literature regarding interpretation of DDA and provide some additional guidance to researchers in the field. The performance of the standardized weights was measured in terms of how well they ordered predictors with varying degrees of between group difference, and what aspects of the data might impact this ordering.

Methods

This Monte Carlo simulation study involved the manipulation of a number of data conditions in order to identify factors influencing the utility of standardized weights for correctly ordering variables based on their relative importance in defining the discriminant function. All analyses were conducted with 2 groups and 6 predictor variables using the SAS software system, version 9.1 (SAS, 2005) PROC DISCRIM. Initially, standardized weights based on both the total and within groups covariance matrices were estimated and retained for further investigation. However, subsequent analysis of the results demonstrated that across all conditions manipulated in this study, the performance of the two types in terms of variable ordering was virtually identical. For this reason, outcomes are reported only for the weights based on the total sample covariance matrices. The manipulated conditions described below were completely crossed with another.

Distribution of the Predictor Variables

The predictor variables were simulated to be normal or non-normal with skewness of 1.75 and kurtosis of 3.75. In order to maintain the desired levels of correlation (described below) among these predictors, the approach for simulating data described by Headrick and Sawilowsky (1999) were employed. These values of skewness and kurtosis were selected because they have been shown to impact the performance of discriminant analysis (Hess, Olejnik, & Huberty, 2001).

Homogeneity of groups' covariance matrices

In addition to the normality of the predictors, a second major assumption underlying DA is the homogeneity of group covariance matrices. Therefore, in order to evaluate the performance of the standardized weights under a range of conditions, the covariance matrices were manipulated to be either equal or unequal. In this study, inequality of covariance matrices was simulated with one group having variances for the predictors that were 5 times larger than that of the other group.

Sample Size

Total sample sizes took four different values across the simulations: 30, 60, 100 and 150. These values correspond to values seen in the applied DA literature (e.g., Glaser, Calhoun, & Petrocelli, 2002; Russell & Cox, 2000; Matters & Burnett, 2003). They represent conditions from small to moderately large samples.

Sample Size Ratio

Three conditions for relative group size were used. In the first condition, the two groups were simulatedwith equal numbers of subjects. In conditions two and three, sample sizes were different such that the larger group had twice the number of subjects as the smaller. In condition two, group 1 had the larger sample size, while in condition three group 2 was the larger. Sample size ratio was completely crossed with covariance matrix equality/inequality. Therefore, in one set of conditions, the larger group had the larger variance while in another the smaller group had the larger variance. In the third combination, the covariances were equal, even as group size ratios were unequal. It was believed that examining the combination of sample size ratio and covariance matrix equality was important to examine because of previous evidence that the interaction of unequal sample sizes and unequal group covariance matrices has an impact on the performance of PDA (Finch & Schneider, 2005).

Group Separation

Separation between the two groups was simulated using Cohen's d, univariate effect size (Cohen, 1988). Table 1 contains the pattern of mean differences for the various combinations of effect sizes. The data were simulated so that group 2 had a mean of 0 and standard deviation of 1 for all of the predictors, while the predictor values for group 1 were generated using the means displayed in Table 1, for each condition respectively. For example, in the 8/0 condition, group 1 had a mean of 0.8 on the first predictor, and means of 0 on the other five, while data for group 2 were generated with means of 0 on all six predictors.

Table 1. Differences (in Cohen's *d*) in Predictor Means between Group 1 and Group 2.

Predictor Variable						Condition Label
X_1	X_2	X_3	X_4	X_5	X_6	
0.5	0	0	0	0	0	5/0
0.8	0	0	0	0	0	8/0
0	0.5	0.5	0.5	0.5	0.5	0/5
0	0.8	0.8	0.8	0.8	0.8	08
3.0	2.5	2.0	1.5	1.0	0.5	5/5
4.8	4.0	3.2	2.4	1.6	0.8	8/8
0.8	0.5	0.5	0.5	0.5	0.5	8/5
0.5	0.8	0.8	0.8	0.8	0.8	5/8

Correlation Between the Predictor Variables

The correlations among the predictors were manipulated at three levels: 0.3, 0.5 and 0.8. In order to maintain these correlations even as the skewed distribution was simulated, the methodology outlined by Headrick and Sawilowsky (1999) was used.

The outcome of interest in this study was the degree to which standardized weights provided correct information regarding the order of importance of predictor variables in terms of group separation. It would be expected that the absolute value of these weights should be larger for those variables associated with greater group separation (Rencher, 1995). Thus, in the context of this study, the weights associated with larger values of Cohen's d should themselves be larger than those weights associated with smaller effect sizes. The specific outcome in this study then, is the proportion of cases across simulation replications in which, for adjacent pairs of variables, the predictor associated with the larger effect size (greater group separation) had the larger standardized weight. In the cases where predictor effect sizes were the same, we anticipate the standardized weights for one of the variables in a pair will be higher than the other roughly half of the time. The appendix contains an S-PLUS program that generated the simulated data for the study. The *rnorm* function in S-PLUS generated 100 random normal data points and output nine variables listed in the data command [data <-c(y,x,z,gain,ypost,xc,xz,xsq,xsqz)]. The post test scores (*Y*) and pre test scores (*X*) were created by adding residual error (*ey* or *ex*) to this random normal variable (*true*). Group assignment (*Z*) was determined based on subtracting a cut score of 20 from the pre test score (1–treatment, 0–comparison). This 10 point treatment gain was added to the post test score (*Y*). Optional *print* and *write* statements are included to either view or save the data in a file.

Results

As mentioned above, results for the total and within groups weights were essentially identical across conditions, therefore the discussion henceforth will focus only on the performance of the total values. In

addition, an examination of results revealed virtually identical outcomes regardless of the sample size ratios simulated. Therefore, in order to limit the length of the manuscript unnecessarily, this variable will also not be included in the following discussion of results. The results are organized by the assumptions of normality and homogeneity of covariance matrices. As stated above, the outcome of interest in this case was the proportion of cases in which the standardized weights correctly reflected the variables' order of importance in terms of group separation. For example, referring to Table 1, in the 5/0 condition, the weight for the first variable should be larger than the weights for the other variables, while the weights for variables 2-6 should be equal (within sampling error) so that no one of these should consistently be larger than the others.

Normal Distribution, Homogeneous Covariance Matrices

Table 2 reflects the results for the case when both assumptions of normality and homogeneous covariance matrices were met, by effect size and correlation among the predictors. Across correlations, when variable 1 had the null effect size and variable 2 did not (05, 08 conditions), the proportion of cases in which variable 2 correctly had the larger standardized weight was greater than 0.90, and increased concomitantly with the correlation value. In contrast, when only the first variable was associated with group separation (80, 50 conditions), the first weight was correctly larger than that of variable 2 at much lower rates. Indeed, for correlations of 0.5 and 0.8, the first weight was correctly larger in less than 30% of the simulation replications.

The proportion of cases in which the standardized weight of the first variable was correctly larger than that of the second in the 88 and 55 conditions, where all of the variables were involved in group separation though to a different degree, was much higher than when only variable 1 differed between groups. Furthermore, as with the 05 and 08 cases, the proportion of replications where the first standardized weight was larger than the second increased along with the correlation among predictors, with the exception of the 55 case for a correlation of 0.8. Finally, when considering the 85 and 58 conditions, the standardized weights were better able to order the variables in the latter case versus the former. In other words, the proportion of cases displaying correct ordering was greater when the second variable had the larger effect size, as opposed to when the first variable had the larger effect size.Note that this outcome follows a very similar pattern to the 05/08 versus 50/80, where variable ordering was correct more frequently in the former than the latter. The proportion of correct ordering outcomes increased with increasing correlation, except for the 85 condition with $r = 0.8$.

When considering the comparisons of the standardized weights for the adjacent pairs in variables 2 through 6, it is important to remember that these variables were all simulated with the same effect size values separating the groups, except for the 88 and 55 conditions. Thus, we would expect them to have very similar standardized weight values across simulation replications. In fact, results for the 80, 50, 08, 05, 85, and 58 conditions revealed that the proportion of times the weights for one of these variables was larger than that of the adjacent one was very close to 0.5 in all cases, indicating that they were comparable in size across replications. Given the similarity of these results in the expected way, the data presented in the tables for variables 2 through 6 only includes rates for the 88 and 55 conditions, where effect size values were not uniform. It is hoped that the tables will more clearly display relevant outcomes that are not obscured by a large number of redundant results.

In general, for both the 88 and 55 conditions it appears that the proportion of cases exhibiting a correct ordering of standardized weights declined somewhat for variable pairs further down the list (e.g., X_3 vs. X_4, X_4 vs. X_5, etc.).For example, in the 88 condition the weight for variable 2 was correctly larger than that of variable 3 at rates comparable to those for the variable 1 versus 2 comparison. In contrast, for the final adjacent pair in the set, variable 5 correctly had a larger standardized weight than variable 6 at lower rates, generally differing by between 0.06 and 0.10 for different values of r. The rate of correct ordering by the standardized weights was higher for larger correlation values, with the exception of the 55 condition with $r = 0.8$.

Table 2. Proportion of cases in which variable ordering is correct based on ldf weights, by effect size and correlation among predictors: ***Normal distribution and homogeneous covariance matrices***.

r	Effect Size	X_1 vs X_2	X_2 vs X_3	X_3 vs X_4	X_4 vs X_5	X_5 vs X_6
0.2	05	0.912				
	08	0.966				
	50	0.578				
	80	0.650				
	58	0.789				
	85	0.809				
	88	0.853	0.860	0.844	0.797	0.761
	55	0.824	0.829	0.797	0.784	0.769
0.5	05	0.945				
	08	0.984				
	50	0.293				
	80	0.227				
	58	0.849				
	85	0.831				
	88	0.885	0.894	0.862	0.827	0.777
	55	0.863	0.866	0.839	0.814	0.803
0.8	05	0.985				
	08	0.999				
	50	0.053				
	80	0.021				
	58	0.940				
	85	0.626				
	88	0.895	0.914	0.883	0.838	0.801
	55	0.603	0.611	0.609	0.588	0.582

Table 3 displays the proportion of correctly ordered variables by effect size and sample size when the assumptions of normality and homogeneity of covariances were met. In general, the pattern of results across effect sizes was very similar to those described above. The proportion of cases correctly ordered for the first 2 variables increased concomitantly with sample size, except for the 50 and 80 conditions. In other words, when only the first variable was simulated to be different between the groups, the proportion of times that the standardized weight for variable 1 was larger than that of variable 2 declined as sample size increased. With respect to the comparisons among the adjacent pairs for variables 2 through 6, the proportion of correctly ordered pairs declined for variables further down the list. In addition, the rate of correct ordering improved with larger sample sizes. Indeed, for a total sample size of 150, the standardized weights were ordered correctly in more than 80% of cases for all adjacent pairs. Even for a sample size of 100, the lowest proportion of accurately ordered pairs was 0.774 for variables 5 and 6 in the 55 condition.

Normal Distribution, Heterogeneous Covariance Matrices

Results for the case where the predictors were simulated to be normally distributed and the covariance matrices between the groups were heterogeneous appear in Tables 4 and 5. Across correlation conditions (Table 4), the proportion of correctly ordered weights was lower than when both assumptions were met (Table 2). The lone exception to this result was for correlations of 0.5 and 0.8 in conjunction with the 50 and 80 effect size conditions, where the proportion correctly ordered was somewhat higher when the covariance matrices were heterogeneous. It should be noted, however, that in general, for the 50 and 80 cases the proportion of correctly ordered weights remained low. The most dramatic reduction in the proportion of correct ordering for the normally distributed heterogeneous covariance case occurred in

Table 3. Proportion of cases in which variable ordering is correct based on ldf weights, by effect size and sample size: ***Normal Distribution and Homogeneous Covariance Matrices***.

N	Effect Size	X_1 vs X_2	X_2 vs X_3	X_3 vs X_4	X_4 vs X_5	X_5 vs X_6
30	05	0.860				
	08	0.946				
	50	0.365				
	80	0.332				
	58	0.753				
	85	0.636				
	88	0.758	0.779	0.750	0.704	0.682
	55	0.688	0.702	0.681	0.655	0.652
60	05	0.951				
	08	0.990				
	50	0.314				
	80	0.288				
	58	0.837				
	85	0.754				
	88	0.869	0.879	0.842	0.801	0.756
	55	0.772	0.781	0.748	0.731	0.717
100	05	0.981				
	08	0.998				
	50	0.279				
	80	0.263				
	58	0.900				
	85	0.833				
	88	0.927	0.931	0.912	0.872	0.813
	55	0.829	0.826	0.808	0.783	0.774
150	05	0.994				
	08	0.999				
	50	0.258				
	80	0.241				
	58	0.936				
	85	0.874				
	88	0.965	0.971	0.947	0.913	0.864
	55	0.863	0.863	0.847	0.830	0.813

the 88 and 55 effect size conditions with r = 0.8. When the data were normally distributed with heterogeneous covariance matrices, the proportion of correctly ordered cases dropped by approximately 0.35 to 0.45 for all adjacent pairs of variables, as compared to the normal homogeneous case.

With respect to the impact of sample size for the normal distribution and heterogeneous covariance condition, results in Table 5 suggest that larger samples did ameliorate the negative impact of heterogeneous covariance matrices for some effect size combinations, but not others. For example, when groups differed on all but the first variable (05, 08), the proportions of correctly ordered standardized weights in Table 5 become very similar to those in Table 3 for samples of 100 and particularly 150. On the other hand, when group separation was isolated in the first variable only (50, 80), the proportion of correctly ordered ldf weights declined with increasing sample size, a pattern also apparent in Table 3. In the other effect size conditions simulated in this study, a larger sample size was associated with improved accuracy in ordering the variables, though the rates did not match those found when both assumptions of normality and homogeneity of variance were satisfied.

Table 4. Proportion of cases in which variable ordering is correct based on ldf weights, by effect size and correlation among predictors: ***Normal Distribution and Heterogeneous Covariance Matrices***.

r	Effect Size	X_1 vs X_2	X_2 vs X_3	X_3 vs X_4	X_4 vs X_5	X_5 vs X_6
0.2	05	0.798				
	08	0.904				
	50	0.522				
	80	0.574				
	58	0.706				
	85	0.708				
	88	0.804	0.803	0.773	0.752	0.736
	55	0.762	0.749	0.741	0.722	0.720
0.5	05	0.843				
	08	0.940				
	50	0.353				
	80	0.304				
	58	0.758				
	85	0.688				
	88	0.838	0.838	0.816	0.785	0.779
	55	0.778	0.772	0.759	0.746	0.741
0.8	05	0.934				
	08	0.979				
	50	0.131				
	80	0.055				
	58	0.854				
	85	0.519				
	88	0.429	0.443	0.453	0.434	0.430
	55	0.203	0.225	0.229	0.236	0.244

Non-Normal Distribution, Homogeneous Covariance Matrices

The third combination of conditions to be examined in this study was the non-normal, homogeneous covariance case. One pattern of results apparent across values of the correlation was that the proportion of correctly ordered weights in the X_1 versus X_2 comparison was higher in the non-normal homogeneous covariance condition than for the normal heterogeneous covariance condition when the first variable was associated with a larger group difference (50, 80, 88, 55). The lone exception to this pattern was the 85 condition, in which the first variable was associated with a large effect while the other variables were associated with a medium effect. Conversely, when the first variable was associated with a null effect size (05, 08) as well as in the 58, 85 cases, the proportion of correct ordering was lower in the non-normal, homogeneous covariance situation. In general, the proportion of correctly ordered standardized weights was lower than when both assumptions were met.

With respect to the adjacent variable comparisons other than X_1 versus X_2, the proportion of correctly ordered weights was somewhat higher earlier in the sequence for the non-normal homogeneous case as compared to the normal heterogeneous data, and somewhat lower for X_4 versus X_5 and X_5 versus X_6. In addition, the sharp decline in accuracy that occurred in the normal heterogeneous case for $r = 0.8$ was not in evidence in the non-normal, homogeneous case. With the exception of the 50 and 80 conditions, the proportion of correctly ordered standardized weights increased with increasing sample sizes in Table 7. In addition, for the 88 and 55 effect size cases, the proportion of correctly ordered weights was comparable or slightly higher in this condition than when both assumptions were met.

Table 5. Proportion of cases in which variable ordering is correct based on ldf weights, by effect size and sample size: ***Normal Distribution and Heterogeneous Covariance Matrices***

N	Effect Size	X_1 vs X_2	X_2 vs X_3	X_3 vs X_4	X_4 vs X_5	X_5 vs X_6
30	05	0.719				
	08	0.844				
	50	0.390				
	80	0.375				
	58	0.670				
	85	0.545				
	88	0.650	0.665	0.648	0.621	0.618
	55	0.570	0.570	0.576	0.568	0.567
60	05	0.837				
	08	0.945				
	50	0.363				
	80	0.317				
	58	0.744				
	85	0.611				
	88	0.710	0.718	0.703	0.675	0.665
	55	0.772	0.621	0.607	0.598	0.595
100	05	0.912				
	08	0.980				
	50	0.312				
	80	0.281				
	58	0.790				
	85	0.688				
	88	0.762	0.757	0.742	0.716	0.709
	55	0.656	0.651	0.635	0.630	0.628
150	05	0.951				
	08	0.994				
	50	0.294				
	80	0.264				
	58	0.862				
	85	0.751				
	88	0.788	0.782	0.764	0.743	0.732
	55	0.680	0.676	0.669	0.655	0.655

Non-normal Distribution, Heterogeneous Covariance Matrices

This combination of conditions represents the situation where neither of the foundational assumptions underlying DA were met. Table 8 reveals that across nearly all conditions the ordering of the standardized weights was correct at markedly lower rates than when both assumptions were met (Table 2). The only exceptions to this pattern were for the 50 and 80 cases, when all group difference was isolated in the first variable only. The pattern of declining accuracy for variables entered later in the equation that was evident in the other distribution and covariance conditions was also apparent when neither assumption was met. In fact, the relative decline in accuracy rates for adjacent pairs further down the sequence was greater in this condition than when both assumptions were met. Larger correlations among the predictors were associated with greater accuracy rates for the 88 and 55 conditions particularly, for the X_1 versus X_2, X_2 versus X_3 and X_3 versus X_4 adjacent pairs. However, for the X_4 versus X_5 and X_5 versus X_6 variable pairs, the proportion of correctly ordered standardized weights actually declined with increasing correlation values.

Table 6: Proportion of cases in which variable ordering is correct based on ldf weights, by effect size and correlation among predictors: ***Non-Normal Distribution and Homogeneous Covariance Matrices***

r	Effect Size	X_1 vs X_2	X_2 vs X_3	X_3 vs X_4	X_4 vs X_5	X_5 vs X_6
0.2	05	0.579				
	08	0.734				
	50	0.527				
	80	0.617				
	58	0.682				
	85	0.683				
	88	0.869	0.836	0.800	0.754	0.637
	55	0.852	0.817	0.775	0.726	0.611
0.5	05	0.571				
	08	0.703				
	50	0.506				
	80	0.575				
	58	0.697				
	85	0.629				
	88	0.907	0.871	0.830	0.733	0.490
	55	0.897	0.853	0.809	0.707	0.472
0.8	05	0.577				
	08	0.693				
	50	0.470				
	80	0.472				
	58	0.763				
	85	0.490				
	88	0.905	0.881	0.812	0.616	0.220
	55	0.665	0.658	0.609	0.514	0.366

The total sample size appears to have been associated with standardized weight ordering accuracy for only some of the effect size combinations when the data were not normally distributed and covariance matrices were not equal between groups. Specifically, from Table 9 when the first variable accounted for more of the group separation (50, 80, 85, 88 and 55 effect size combinations) the proportion of correctly ordered weights for the X_1 versus X_2 comparison increased concomitantly with sample size. This increase in accuracy was most notable in the 88 and 55 cases. For adjacent pairs other than X_1 and X_2, there was a clear positive relationship between sample size and weight ordering accuracy for the X_2 versus X_3 and X_3 versus X_4 comparisons. On the other hand, for the last two pairs in the sequence, there appears not to have been this positive relationship between sample size and the accuracy rate.

The goal of this Monte Carlo study was to examine the potential utility of standardized weights for ordering predictor variables in terms of their relative importance in defining a significant discriminant function. Prior simulation research has found that other methods for characterizing group separation in DDA, such as the use of SC's, may be less than optimal in many situations. Thus, the current research was designed to ascertain how effective an alternative the standardized weights might be for this purpose. The study conditions were selected so as to replicate those in earlier studies that focused on SC's, and the outcome of interest was the proportion of cases in which the weights correctly ordered the variables in terms of their relative importance in separating two groups.

Table 7: Proportion of cases in which variable ordering is correct based on ldf weights, by effect size combination and sample size: ***Non-Normal Distribution and Homogeneous Covariance Matrices***

N	Effect size	X_1 vs X_2	X_2 vs X_3	X_3 vs X_4	X_4 vs X_5	X_5 vs X_6
30	05	0.511				
	08	0.570				
	50	0.502				
	80	0.534				
	58	0.609				
	85	0.532				
	88	0.780	0.744	0.694	0.617	0.468
	55	0.711	0.681	0.644	0.588	0.488
60	05	0.555				
	08	0.673				
	50	0.501				
	80	0.538				
	58	0.687				
	85	0.578				
	88	0.879	0.836	0.782	0.681	0.447
	55	0.792	0.758	0.707	0.635	0.483
100	05	0.599				
	08	0.772				
	50	0.502				
	80	0.572				
	58	0.762				
	85	0.620				
	88	0.943	0.916	0.855	0.736	0.440
	55	0.840	0.814	0.763	0.672	0.483
150	05	0.638				
	08	0.825				
	50	0.500				
	80	0.576				
	58	0.798				
	85	0.671				
	88	0.974	0.954	0.915	0.772	0.441
	55	0.876	0.851	0.810	0.699	0.484

Discussion

The results described above indicated that under some conditions, the standardized weights did indeed provide an accurate ordering of the predictor variables, particularly when both the assumptions of normality and homogeneity of covariance matrices were met. These accuracy rates were frequently over 90% for samples of 100 and 150 subjects. Furthermore, the ordering accuracy rates for all adjacent pairs improved when the correlations among the predictors increased in several of the conditions simulated here. The major exception to these positive results when both assumptions were met occurred when group separation was only present for the first predictor variable. In this case, the accuracy rates were much lower than for the other conditions, and they declined with increasing correlations among the variables. In other words, when the group difference was truly univariate in nature and centered in the first variable, the standardized weight for the second variable was frequently (incorrectly) larger than that of the first. Finally, the accuracy of the standardized weight ordering approach was somewhat higher for variable pairs earlier in the sequence, even though the relative difference in group separation later in the

Table 8. Proportion of cases in which variable ordering is correct based on ldf weights, by effect size and correlation among predictors: Non-Normal distribution and heterogeneous covariance matrices

r	Effect Size	X_1 vs X_2	X_2 vs X_3	X_3 vs X_4	X_4 vs X_5	X_5 vs X_6
0.2	05	0.503				
	08	0.484				
	50	0.547				
	80	0.587				
	58	0.510				
	85	0.518				
	88	0.759	0.714	0.648	0.565	0.439
	55	0.648	0.596	0.557	0.525	0.444
0.5	05	0.487				
	08	0.445				
	50	0.569				
	80	0.634				
	58	0.509				
	85	0.527				
	88	0.821	0.757	0.679	0.527	0.303
	55	0.708	0.640	0.576	0.473	0.371
0.8	05	0.488				
	08	0.440				
	50	0.597				
	80	0.695				
	58	0.479				
	85	0.569				
	88	0.888	0.830	0.698	0.430	0.167
	55	0.793	0.726	0.596	0.433	0.253

sequence was identical. For example, Table 1 shows that the difference between group means for variable 2 was simulated to be 4.0 in the 88 effect size case, while the difference for variable 3 was simulated to be 3.2. Thus the difference in conditions was 0.8 (4.0-3.2). The difference between group means for variable 4 was simulated to be 2.4, which was 0.8 units different from the group separation for variable 3. However, the proportion of correctly ordered weights for variable 2 versus variable 3 was greater than that for variable 3 versus variable 4 across correlation conditions. A similar pattern was evident for the other adjacent variable pairs further down the sequence.

In general, the results of this study demonstrated that when the assumptions of normality and/or homogeneity of covariance matrices were not met, the standardized weights were less accurate in ordering predictor variables based on their relative importance in group separation. The performance of these weights was generally most degraded when neither assumption was met. The lone exception to this last pattern occurred when the predictors were normally distributed but the covariance matrices were unequal and the correlation among the predictors was 0.8. In this case, the ordering accuracy rates were well below 50% for both the 88 and 55 effect size conditions. Under most conditions where one or both of these assumptions were unmet, larger sample sizes served to mitigate problems with ordering accuracy to some extent, though rarely did accuracy match that when both assumptions were met. The positive impact of increased sample sizes was particularly evident when the data were non-normal. Indeed, in the 88 and 55 effect size conditions, the accuracy rates were comparable (or nearly so) to the normal, homogeneous covariance case for both of the non-normal situations when the sample size was 150. It should also be noted that when the first variable was not associated with group separation (08, 05) the accuracy rates in the normal distribution, heterogeneous covariance condition were higher than when the data were not normally distributed, and for samples of 100 and 150 were above 0.9.

Implications for Practice

Some authors (e.g., Rencher, 1995) have recommended that researchers using DDA to differentiate two or more groups in the multivariate case consider relying on these standardized weights to characterize the nature of the significant discriminant functions. Rencher (1992) argued that they are superior to other tools, such as SC's, because they incorporate information about all of the variables in the analysis, rather than simply reproducing univariate analyses. The results of this study appear to support the potential utility of these standardized weights for characterizing multivariate group differences in some situations, but not others. Following are some potential implications for practice based on results discussed above. It should be noted that guidelines for what would be acceptable performance are not available. Ideally, of course, the rates of correct variable ordering would be 100%, though such a perfect outcome would be unlikely for any statistical procedure. Rather than select an arbitrary cut off for what is acceptable performance, we have elected in this manuscript to discuss the rates in relative terms and allow readers to make their own judgments regarding the acceptability (or not) of the standardized weights' performance.

First of all, it does appear that when the assumptions of normality and group homogeneity of covariance matrices are both satisfied, variables are accurately ordered in terms of relative contribution to group separation at rates above 80% when the sample size is 100 or greater and the group differences are multivariate in nature (all effect size conditions except for 80 and 50). Indeed, when the sample size was as least 60 and all the variables were associated with group separation, the standardized weights would accurately order variables 1 and 2 in importance more than 80% of the time, except when the second variable was associated with a moderate effect and the first was associated with a moderate or large effect (85, 55 conditions).

While performance of the weights in variable ordering was often relatively goodwhen the groups were separated on multiple predictors (and the foundational assumptions were met), in cases where the groups only differed on one variable (the first in the sequence in this study), theydid not accurately reflect this fact very well, regardless of sample size. This problem was more acute when the predictor variables were more highly correlated. Therefore, researchers using DDA should carefully consider the variables that they have selected as predictors so that any significant group differences not beunivariate in nature. Furthermore, if results of the analysis appear to indicate that the groups differ on only one variable, the researcher should be very careful when interpreting variable ordering with these standardized weights.

When the predictor variables do not conform to the assumptions of normality and homogeneity of covariance matrices, researchers should also exercise caution when using standardized weights to interpret discriminant functions. Results of this study suggest that when the predictor variables are not normally distributed and/or the group covariance matrices are not equal, the weights may frequently order the variables incorrectly in terms of their relative importance, particularly when both assumptions are violated simultaneously. Therefore, researchers considering the use of these weights for characterizing the nature of significant group separation should be very careful to check these assumptions. If they do not hold, the weights may not be appropriate for ordering the variables. It is important to note that larger overall sample sizes do not fully ameliorate this problem.

A fourth implication of these results is that the correlations among the predictor variables have an impact on the performance of standardized weights when the assumptions of normality and homogeneity of covariances are met. In general, higher correlations among the predictors were typically associated with more accurate ordering based on the standardized weights. The lone exception to this outcome occurred when only the first variable was associated with group difference, in which case higher correlations resulted in the weight of the second variable (not different between groups) being larger (incorrectly) than that of the first, at very high rates. Researchers considering the use of standardized weights for interpreting DDA thus need to be cognizant of these correlations. If they select a number of variables that have relatively low correlations, they may have more difficulty in correctly identifying which of these is most associated with the significant discriminant function, and the associated group differences. It is also interesting to consider this result in light of Rencher's (1992) argument in favor of using standardized weights: namely that they account for the presence of the other predictors in the model. The fact that performance generally improved with higher correlations appears to validate this earlier observation.

Finally, when compared with results of earlier simulation research examining the SC's as a tool for interpreting discriminant functions, the standardized weights appear to perform favorably. Finch (2007) reported very high rates (often in excess of 0.5) of incorrect identification of "important" variables using

these SC's. In addition, under several data conditions similar to those included in this study, rates of correct identification of such "important" variables were not higher than those reported here for the standardized weights. Therefore, given the high Type I error rates for the SC's, along with the comparable power, it would appear that the standardized weights may prove to be a worthwhile alternative for interpreting significant discriminant functions.

Limitations and Directions for Future Research

Future studies should be designed to improve on the current research. For example, results described in this manuscript are limited to the two groups case. Thus, one logical next step in this area is to examine the utility of standardized weights for differentiating among more than two groups. By including multiple groups, interpretation of more than one significant discriminant function would also be possible.

A second area for future research is the examination of the performance of standardized weights for a different set of effect size combinations. In the current study, most of the differences among the predictors with respect to group separation were between variable 1 and the others. With the exception of the 88 and 55 conditions, variables 2 through 6 were associated with the same effect size difference between the groups. Future studies should use a different variety of such group differences in order to provide a more complete understanding of the effectiveness of the weights for ordering the predictor variables.

Future studies in this area should also examine a different set of non-normal distributions for the predictors. While this is the first study in this area to use non-normal data, generalizations of the results herein are limited to those non-normal cases where the predictors have skewness of 1.75 and kurtosis of 3.75. For example, some research has shown that a related statistical analysis, Multivariate Analysis of Variance (MANOVA), is impacted by variables with truncated tails (e.g., Finch, 2005). Thus, it seems reasonable that DDA, which is based upon the same multivariate linear model, might also experience problems with such a distribution.

References

Cohen, J. (1988). *Statistical power analysis for the social sciences* (2nd. ed.). Hillsdale, NJ: Erlbaum.

Dalgleish, L.I. (1994). Discriminant Analysis: Statistical inference using the Jackknife and Bootstrap procedures. *Psychological Bulletin, 116,* 498-508.

Finch, W.H. (2007). *Identification of variables associated with group separation in Descriptive Discriminant Analysis: Comparison of methods for interpreting Structure Coefficients*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Finch, W.H. & Schneider, M.K. (2005). 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,* 240-257.

Glaser, B.A., Calhoun, G.B., & Petrocelli, J.V. (2002). Personality characteristics of male juvenile offenders by adjudicated offenses as indicated by the MMPI-A. *Criminal Justice Behavior, 29,* 183-201.

Headrick, T.C. & Sawliowsky, S.S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika, 64,* 25-35.

Hess, B., Olejnik, S., &Huberty, C.J. (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,* 909-936.

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

Huberty, C. J. & Olejnik, S. (2006). *Applied Manova and Discriminant Analysis.* New York: Wiley.

Huberty, C.J. & Wisenbaker, J.M. (1992). Variable importance in multivariate group comparisons. *Journal of Educational Statistics, 17,* 75-91.

McClachlan, G.J. (1992). *Discriminant Analysis and Pattern Recognition.* New York: Wiley.

Matters, G. & Burnett, P.C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement, 63,* 239-256.

Meshbane, A. & Morris, J.D. (1996). *Predictive Discriminant Analysis versus Logistic Regression in two-group classification problems*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

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

Rencher, A.C. (1995). *Methods of Multivariate Analysis*. New York: Wiley.

Rencher, A.C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician, 46*, 217-225.

Russell, W.D. & Cox, R.H. (2000). Construct validity of the Anxiety Rating Scale-2 with individual sport athletes. *Journal of Sports Behavior, 23(4)*, 379-388.

SAS Version 9.1, (2005). Cary, NC: The SAS Institute.

Stevens, J. (2000). *Applied Multivariate Statistics for the Social Sciences*. Mahwah, NJ: Lawrence Erlbaum Associates, publishers.

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

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

Impact of Rater Disagreement on Chance-Corrected Inter-Rater Agreement Indices with Equal and Unequal Marginal Proportions

David A. Walker
Northern Illinois University

This study examined the effect that equal free row and column marginal proportions, unequal free row and column marginal proportions, and the magnitude of rater disagreement had on eight agreement indices. In condition 1, when there were equal free row and column marginal proportions with no rater disagreement present, seven of the eight indices of agreement yielded very comparable results. In condition 2, when there were unequal free row and column marginal proportions and rater disagreement was ≤ .10, five of the eight indices of agreement tended to produce similar results. In conditions 3 and 4, when the marginals were not homogeneous and the amount of rater disagreement was > .10, there were three instances each of over-estimation and under-estimation. Thus, as cells B and C became less homogeneous, all of the inter-rater agreement indices studied, except for Cohen and Dice, were influenced via under- or over-estimation once rater disagreement was > .10. If rater disagreement was ≤ .10, 5 out of the 8 indices studied were not influenced by some degree of marginal heterogeneity.

In social science research, inter-rater agreement indices of categorical data for two raters have been studied extensively, and their strengths and weaknesses in various methodological situations reviewed in contexts such as classroom observations, political polling, psychological analysis, and content analysis (Bennett, Alpert, & Goldstein, 1954; Krippendorff, 2004; Riffe & Freitag, 1997; Zwick, 1988). Inter-rater agreement is conducted to verify that rater agreement exceeds, or does not, chance levels of agreement. The range of rater agreement is from -1.00 to +1.00, with +1.00 as total agreement, 0 as not better than chance that the raters would agree, and negative results indicate agreement worse than expected by chance due to random or systematic differences between raters such as rater bias or coding errors (Kassarjian, 1977; Linn & Gronlund, 2000; Sim & Wright, 2005).

In the literature pertaining to inter-rater agreement, various indices used with two raters and binary data emerge. All of these indices use a 2 x 2 agreement matrix, where the main diagonal (i.e., cells A and D) indicates the agreement level between the raters as either 00 or 11 and the off diagonal (i.e., cells B and C) indicates the level of disagreement between the raters as either 10 or 01.

		Rater 2		
		0	**1**	**Total**
Rater 1	**0**	Cell A	Cell B	$\mathbf{p_1}$
	1	Cell C	Cell D	$\mathbf{q_1}$
	Total	$\mathbf{p_2}$	$\mathbf{q_2}$	**n**

Figure 1. A 2 x 2 Matrix Configuration.

There are numerous indices for inter-rater agreement corrected for chance that can be applied to 2 x 2 tables with categorical data. However, in the scholarly literature (Fleiss, 1975; Hertzberg, Xu, & Haber, 2006; Krippendorff, 2004; Rae, 1988; Sirotnik, 1981; Übersax, 1987; Zwick, 1988), the following eight measures of agreement have been noted as common indices used and can be defined as "… proposed for categorical response data where such response is the assignment of the subject to one of κ mutually exclusive and exhaustive categories. [and] as a measure of agreement between multiple observations of a single subject" (Kraemer, 1979, p. 461).

The first chance corrected index for inter-rater agreement using a 2 x 2 table was proposed by Bennett et al. (1954) as Bennett's *S* coefficient, which requires the assumption of uniform marginals, where:

$$S = \frac{k}{k-1}(P_o - \frac{1}{k}) \quad (1)$$

where, P_o = observed agreement, where P_o = A + D; A = count from cell A, D = count from cell D; and k = number of response categories. Scott (1955) proposed Scott's pi coefficient or π, which requires the assumption of homogeneous marginals for the raters, where:

$$\pi = \frac{P_o - P_e}{1 - P_e} \; . \quad (2)$$

Note: Fleiss' intraclass correlation coefficient (1975) in a 2 x 2 situation is the same formula as Scott's π, with the assumption of equally distributed marginals where,

$P_e = \sum_{i=1}^{k} \frac{(n_{i.} + n_{.i})^2}{2}$, is expected percentage of agreement based on chance, k = number of response categories. $n_{i.}$ = observed row marginals for response i for rater 1, and $n_{.i}$ = observed column marginals for response i for rater 2.

Cohen (1960) proposed Cohen's kappa coefficient or κ (1960), but did not have an assumption related to equally-distributed marginals, yet did assume that "… N objects categorized are independent; the assigners operate independently; and the categories are independent, mutually exclusive, and exhaustive" (Brennan & Prediger, 1981, p. 688).

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

where, A = count from cell A, B = count from cell B, C = count from cell c, D = count from cell D,
$P_o = (A + D)/N$ is the observed agreement, N = number of observations, and
$P_e = [(A + B)*(A + C)+(C + D)*(B + D)]/N^2$ is the expected percentage of agreement based on chance.

Armitage, Blendis, and Smyllie (1966) proposed the standard deviation index, which is very similar to κ with no distributional assumption, but with the same assumptions of independence. Equations 5 to 7 do not have a distributional assumption, but also have the same independence assumptions as κ.

$$SD = \frac{(AD - BC)(p_1q_1 + p_2q_2)}{2p_1q_1p_2q_2} \; ; \tag{4}$$

where $p_1 = (A+B)$, $p_2 = (A+C)$, $q_1 = (C+D)$, and $q_2 = (B+D)$.

Maxwell and Pilliner (1968) proposed r_{MP}:

$$r_{MP} = \frac{2(AD - BC)}{(p_1q_1 + p_2q_2)} \; ; \tag{5}$$

The phi coefficient (as cited in Fleiss, 1975) was proposed as an inter-rater index:

$$\Phi = \frac{(AD - BC)}{\sqrt{p_1q_1p_2q_2}} \; ; \tag{6}$$

The Dice index (as cited in Fleiss, 1975) was proposed as a measure of inter-rater agreement:

$$S_D = \frac{2(AD - BC)}{(p_1q_2 + p_2q_1)} \; ; \tag{7}$$

Methods

The data for the subsequent situations tested on each of the inter-rater agreement indices were derived from an SPSS (Statistical Package for the Social Sciences v. 15.0) program written by the author. Each of the 2 x 2 situations used binary data; there were no missing data; each situation had free, homogeneous or heterogeneous marginals (i.e., "… a margin is 'free' whenever the marginal proportions are not known to the assigner beforehand" (Brennan & Prediger, 1981, p. 690); and each situation had either no rater disagreement, rater disagreement ≤.10, rater disagreement >.10 but ≤.20, or rater disagreement >.20. Rater disagreement was determined from the following formula presented in Sim and Wright (2005):

$$\text{Rater Disagreement} = \left| B - C \right| / N \tag{8}$$

An SPSS bootstrap program created by the author was used. The bootstrap is a resampling method where the sampling properties of a statistic, in this instance the inter-rater agreement indices, are derived by recomputing their value for artificial samples. Thus, the sample data from this study served as pseudo-populations and 20,000 random samples with replacement were drawn from these full samples. Twenty

thousand iterations were used as an established threshold where all of the four cases in this study had convergence. Once the bootstrap method was repeated 20,000 times on each of the four cases, a distribution of bootstrapped estimates for the kappa-related indices emerged, where the mean value (i.e., κ_{Boot}) of each bootstrapped distribution was the estimate for each of the four case's population κ value. Further, the bootstrap was employed as a method for estimating generalization error, which, in turn, was used to form 95% confidence intervals around the κ_{Boot}.

Using Monte Carlo generated data, the purpose of this research was to examine the possible effect that equal free row and column marginal proportions (EM), unequal free row and column marginal proportions (UM), and the magnitude of rater disagreement had on the agreement indices under study. As James (1983, p. 651) noted 25 years ago, "Much less attention seems to have been paid to the analysis of nonagreements…"

Results

Table 1 shows the kappa-related statistics in each of the four conditions by sample sizes of 10, 20, 50, and 75 typically found in educational research (Claudy, 1972; Huberty & Mourad, 1980). The bootstrap results from Table 1 denote which of the eight indices were outside of the confidence intervals established as thresholds for each case pertaining to under-estimation or over-estimation of inter-rater agreement given the circumstances of homogeneous or heterogeneous marginals and no rater disagreement to some level of disagreement.

The results found in Table 1 indicated that in the first case, when the marginals were homogeneous (i.e., verified via a McNemar's Test based on difference in the marginal probability distribution between observations in a 2 x 2 matrix, where Ho: p1. = p.1 and H1: p1. ≠ p.1) and there was no rater disagreement present, seven of the eight indices showed no under- or over-estimation of inter-rater agreement, which was an expected assumption in this situation with all of the kappa-like formulas (note: the lone exception of over-estimation was found with the Bennett index).

In the second case, when the marginals were not homogeneous and the amount of rater disagreement was ≤ .10, there was one instance of over-estimation with the Bennett index and two occurrences of under-estimation found with the Fleiss and Scott indices. In the third case, when the marginals were not homogeneous and the amount of rater disagreement was > .10 but ≤ .20, there were three instance of over-estimation with the Phi, Maxwell-Pilliner, and Armitage et al. indices, and three occurrences of under-estimation found with the Bennett, Fleiss, and Scott indices. Finally, in the fourth case, when the marginals were not homogeneous and the amount of rater disagreement was > .20, all of the same indices from case 3 that had over- or under-estimation problems repeated in case 4. That is, there was noticeable over-estimation associated with Phi, Maxwell-Pilliner, and Armitage et al., and evident occurrences of under-estimation found with Bennett, Fleiss, and Scott.

Discussion

Thus, given the similar assumptions affiliated with kappa-like indices of agreement, when there were equal free row and column marginal proportions with no rater disagreement present, seven of the eight indices of agreement yielded the same results. This outcome was expected based on the assumption of marginal homogeneity for many of the kappa-like measures. When there were unequal free row and column marginal proportions and rater disagreement is ≤ .10, five of the eight indices of agreement tended to produce similar results, with two of the three deviant indices very close to the established confidence interval (e.g., Scott and Fleiss within .001).

When the marginals were not homogeneous and the magnitude of rater disagreement was > .10, cases 3 and 4 showed a trend in indices that succumbed to over- and under-estimation. That is, when rater disagreement was evident (i.e., > .10), there should be some caution used when applying the Phi, Maxwell-Pilliner, and Armitage et al. indices in a 2 x 2 situation due to their tendency to over-estimate chance-corrected agreement, and some prudence employed when using the Bennett, Fleiss, and Scott indices due to their propensity to under-estimate chance-corrected agreement when compared to other commonly-used indices of agreement.

Implications and Conclusions

Overall, the data trends indicated that the Bennett index either over- or under-estimated chance-corrected agreement in a 2 x 2 situation in all four cases studied regardless of the presence, or lack thereof, of rater disagreement. The Fleiss and Scott indices under-estimated in three of the four cases (i.e., contingent upon some level of rater disagreement).

Table 1. Measures of Agreement Bootstrap Results

Sample	**N = 10**	**N = 20**	**N = 50**	**N = 75**
Cell Counts	A = 2, B = 2, C = 2, D = 4	A = 8, B = 3, C = 4, D = 5	A = 19, B = 4, C = 12, D = 15	A = 30, B = 3, C = 21, D = 21
Rater Disagreement	**0**	**≤ .10**	**> .10 ≤ .20**	**> .20**
Agreement Index				
Cohen	.167	.286	.372	.387
Maxwell-Pilliner	.167	.287	.392*	.435*
Scott	.167	.284*	.356*	.351*
Fleiss	.167	.284*	.356*	.351*
Armitage et al.	.167	.287	.392*	.436*
Dice	.167	.286	.372	.387
Phi	.167	.287	.392*	.435*
Bennett	.200*	.300*	.360*	.360*
Bootstrap				
Mean: κ_{Boot}	.171	.288	.374	.393
Standard Deviation	.004	.002	.005	.013
95% Confidence Interval	(.167, .179)	(.285, .292)	(.364, .385)	(.368, .419)

***** = Outside of confidence interval range

As seen in Table 2, the Bennett, Scott, and Fleiss indices, which all adhered to the assumption of homogeneous marginals, preformed the poorest when any level of rater disagreement was present and, thus, their use in situations of disagreement is not recommended.

The Phi, Maxwell-Pilliner, and Armitage et al. indices over-estimated in two of the four cases, particularly when rater disagreement > .10, Therefore, the recommendation found in Table 2 is to employ these indices when rater disagreement is ≤ .10. Cohen and Dice were the only indices that did not manifest any penchant to over- or under-estimate chance-corrected agreement when confronted with rater disagreement and are recommended as reliable measures in all conditions tested.

Table 2. Recommendations for the Use of Agreement Indices per Level of Rater Disagreement

Rater Disagreement	**0**	**≤ .10**	**> .10 ≤ .20**	**> .20**
Agreement Index				
Cohen	*	*	**	**
Maxwell-Pilliner	*	*	NR	NR
Scott	*	NR	NR	NR
Fleiss	*	NR	NR	NR
Armitage et al.	*	*	NR	NR
Dice	*	*	**	**
Phi	*	*	NR	NR
Bennett	NR	NR	NR	NR

NR = Not Recommend for use
***** = Use in conditions of rater disagreement **≤ .10**
****** = Use in conditions of rater disagreement **> .10**

An implication affiliated with the current study may be seen in the area of contributing to the base in the scholarly literature, where this is one of very few studies (cf. Whitehurst, 1984) that has looked at the magnitude that rater disagreement has on various inter-rater agreement indices. As Zwick (1988) noted about the degree that marginal homogeneity may play in inter-rater agreement indices, "Rather than ignoring marginal disagreement or attempting to correct for it, researchers should be studying it to determine whether it reflects important rater differences or merely random error" (p. 377). A second implication is that this research provides guidelines concerning which of the frequently used measures of agreement would be plausible options to employ when a level of rater disagreement is present.

References

Armitage, P., Blendis, L. M., & Smyllie, H. C. (1966). The measurement of observer disagreement in the recording of signs. *Journal of the Royal Statistical Society, Series A, 129*, 98-109.

Bennett, E. M., Alpert, R., & Goldstein, A. C. (1954). Communications through limited response questioning. *Public Opinion Quarterly, 18*, 303-308.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687-699.

Claudy, J. G. (1972). A comparison of five variable weighting procedures. *Educational and Psychological Measurement, 32*, 311-322.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Fleiss, J. L. (1975). Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31*, 651-659.

Hertzberg, V. S., Xu, F., & Haber, M. (2006). Restricted quasi-independent model resolves paradoxical behaviors of Cohen's kappa. *Journal of Modern Applied Statistical Methods*, *5* (2), 417-431.

Huberty, C. J., & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement, 40*, 101-112.

James, I. R. (1983). Analysis of nonagreements among multiple raters. *Biometrics, 39*, 651-657.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research, 4*, 8-18.

Kraemer, H. C. (1979). Ramifications of a population model for κ as a coefficient of reliability. *Psychometrika, 44*, 461-471.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*, 411-433.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, NJ: Prentice-Hall.

Maxwell, A. E., & Pilliner, A. E. G. (1968). Deriving coefficients of reliability and agreement for ratings. *British Journal of Mathematical and Statistical Psychology, 21*, 105-116.

Rae, G. (1988). The equivalence of multiple rater kappa statistics and intraclass correlation coefficients. *Educational and Psychological Measurement, 48*, 367-374.

Riffe, D., & Freitag, A. A. (1997). A content analysis of content analyses: Twenty-five years of *Journalism Quarterly*. *Journalism & Mass Communication Quarterly, 74*, 873-882.

Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321-325.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*, 257-268.

Sirotnik, K. A. (1981). Assessing attitudinal congruence: A case for absolute (as well as relative) indices. *Journal of Educational Measurement, 18*, 205-212.

Übersax, J. S. (1987). Diversity of decision making models and the measurement of interrater agreement. *Psychological Bulletin, 101*, 140-146.

Whitehurst, G. J. (1984). Interrater agreement for journal manuscript reviews. *American Psychologist, 39*, 22-28.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103*, 374-378.

Send correspondence to: David A. Walker
Northern Illinois University
Email: dawalker@niu.edu

Multiple Linear Regression Viewpoints

Information for Contributors

Multiple Linear Regression Viewpoints (*MLRV*) is a publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression: The General Linear Model (MLR/GLM SIG). It is published twice a year to facilitate communication among professionals who focus their research on the theory, application, or teaching of multiple linear regression models and/or the general linear model. Manuscripts submitted to *MLRV* should conform to the language, style, and format of the Publication Manual of the American Psychological Association (5th ed., 2001). Three copies (two blind) of a doubled spaced manuscript (including equations, footnotes, quotes, and references) of approximately 25 pages in length, a 100 word abstract, and an IBM formatted diskette with the manuscript formatted in WordPerfect or Word should be submitted to one of the editors listed below.

Mathematical and Greek symbols should be clear and concise. All figures and diagrams must be photocopy-ready for publication. Manuscripts will be anonymously peer reviewed by two editorial board members. Author identifying information should appear on the title page of only one submitted manuscript. The review process will take approximately 3 months. A letter acknowledging receipt of the manuscript will be sent to the first author, and upon review, a letter indicating the peer review decision will be sent to the first author. Potential authors are encouraged to contact the editors to discuss ideas for contributions or determine if their manuscript is suitable for publication in *MLRV*.

EDITOR INFORMATION

Randall E. Schumacker, Editor *MLRV*
College of Education
P.O. Box 870231
Carmichael Hall 316
The University of Alabama
Tuscaloosa, AL 35487-0231
(205) 348-6062
rschumacker@ua.edu

T. Mark Beasley, Associate Editor
Department of Biostatistics
School of Public Health
309E Ryals Public Health Bldg.
University of Alabama, Birmingham
Birmingham, AL 35294
(205) 975-4957
mbeasley@uab.edu

ORDER INFORMATION

Cynthia Campbell, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854
ccampbell@niu.edu

Check out our website at: http://mlrv.ua.edu/

POSTMASTER: Send address changes to:
Cynthia Campbell, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the
AERA Special Interest Group on Multiple Linear Regression: General Linear Model
through the **University of Alabama at Birmingham** and the **Dallas Independent School District**.