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 30 • Number 1 • Spring 2004

Table of Contents

Comparison of Logistic Regression, Linear Probability, and Third-Degree Polynomial Models: Which Should a Researcher Use?

Isadore Newman University of Akron
Russell Brown University of Akron
John W. Fraas Ashland University

The purpose of this study is to present a comparison of three types of regression models that could be used to analyze dichotomous criterion variables under three different data structures, and to discuss the implications of the results of those comparisons. The three types of models used were (a) logistic regression, (b) linear ordinary least squares, and (c) polynomial ordinary least squares models.

Logistic regression is commonly used in medical literature as a means to account for the variance in a binary (or categorical) dependent variable (King & Ryan, 2002), and its use is growing in the social sciences literature as well. Peng, So, Stage, and St. John (2002) reported that "research using logistic regression has been published with increasing frequency in three higher education journals: *Research in Higher Education*, *The Review of Higher Education*, and *The Journal of Higher Education*" (p. 259). This trend has corresponded with the increased availability of computer software that provides the option to analyze data using logistic regression (Peng, Lee, & Ingersoll, 2002). While there has been an increase in the use of this method, its use has been accompanied by "great variation in the presentation and interpretation of results in these publications, which can make it difficult for readers to understand and compare the results across articles" (Peng, So, Stage, & St. John, 2002, p. 259).

The popularity of logistic regression has grown, in part, due to proponents who have suggested that it is a more appropriate alternative to ordinary least square (OLS) linear regression or discriminant analysis for modeling categorical (dichotomous) dependent variables. With a dichotomous dependent variable, all of the observed dependent data points will fall on one of two horizontal lines that are parallel, which is a difficult condition to model with the single straight line produced by an OLS linear model. Peng, Lee, and Ingersoll (2002) suggested a potential solution to this problem via plotting the calculated means of the dependent variables for categories of the independent variable. Such a plot takes a sigmoid shape, which Peng, Lee and Ingersoll rightly point out, has extremes that do not follow a linear trend.

Perhaps, it should be no surprise then that a linear fit to such data has obvious limitations. In addition, the errors in this type of model are not normally distributed and are not constant across the range of the data. Finally, OLS models produce values that are above (greater than 1) and below (less than 0) the range of the observed levels of the dependent variable. This problem has been partially addressed by constraining the results of the predicted probabilities to a logical range, but this comes at the expense of treating values above and below the range as perfectly representative of the end points (i.e., 100% likely for points above 1 and 0% likely for points below 0).

The growth in the use of logistic methods is predicated on the reported "superiority of logistic regression over OLS models" (Peng, So, Stage, & St. John, 2002, p. 260) as a means to overcome the "limitations of ordinary least squares (OLS) regression in handling dichotomous outcomes" (Peng & So, 2002, p. 31). The logistic model is as follows:

$$\ln[p/(1-p)] = \alpha + \beta X + e \ ,$$

where: p = probability that the event Y occurs, α = the Y intercept, β = the regression coefficient, and e =error. The natural log transformation of the odds ratio is necessary to make this relationship linear. The most obvious advantage of this model is that it constrains the predicted values of probability to the logical range of 0 to 1, which overcomes an obvious limitation of the OLS model. Additionally, the model does not require "data that are drawn from a multivariate normal distribution with equal variances and covariances for all the variables" (Peng & So, 2002); and, therefore, it has less restrictive assumptions than either OLS or linear discriminant function analysis.

Brown and Newman (2002) examined methods for modeling data that conformed to a known function or shape and found that, in some instances, polynomial modeling could be superior to modeling based upon the known function (e.g., cosine). A sigmoid curve could be modeled using a polynomial function that accounted for the two inflection points in the curve:

$$Y = a_0U + a_1X + a_2X^2 + a_3X^3 + e$$

A polynomial model of this nature would potentially better fit the shape of the distribution at its extremes; and, if consistent with the research of Brown and Newman, it would be better able to fit the data if the data deviated from the sigmoid shape.

While many have expressed concerns about the limitations of OLS when predicting dichotomous outcomes, Pohlmann and Leitner (2000) found very similar results when they compared OLS and logistic regression methods. Indeed, they found identical conclusions in regard to significance testing, and they found the predicted values for both modeling methods to be quite similar. Pohlmann and Leitner did, however, find a slight advantage in accuracy of the predicted values. The similarity in results is striking given that Pohlmann and Leitner used a linear model (OLS) as the basis of the comparison.

Method

The method for the present study was an extension of the method described by Pohlmann and Leitner (2000). Common artificial data sets were used to compare the three methods (i.e., linear, polynomial, and logistic regression) in terms of the effects of known changes in the distribution of the scores on the results of the analysis.

Each of the distributions of the independent variable was created to appear like a group of 200 intelligence test scores. In the first case, which is shown in Figure 2, the scores, which reflected a bimodal distribution (Modes = 94 and 103, $M = 98.5$, $s = 3.154$, Range = 15), were highly correlated with the dependent variable ($r = .837$). In the second case, which is shown in Figure 3, the scores were more normally distributed ($M = 97.9$, $s = 5.636$, Range = 26) and moderately correlated ($r = .445$) with the dependent variable. In scores in the third case (which is shown in Figure 4), were normally distributed ($M = 97.2$, $s = 5.148$, Range = 31) and slightly correlated with the dependent variable ($r = .150$).

Linear, polynomial, and logistic regression analyses were performed on each of the three data sets, and the three methods were compared in terms of (a) the goodness-of-fit values and the statistical tests of those values, (b) the correlations of the predicted probabilities, (c) the mean square error values, and (d) the accuracy of the classifications of group membership.

Results

The three sets of data were analyzed by using a linear OLS model, a polynomial OLS model, and a logistic model. In order to facilitate the retention of all the variables in the polynomial model (i.e., linear, squared, and cubic variables), the scores were centered before these variables were generated. The results of these analyses are contained in Table 1.

Goodness-of-Fit Values

The first comparison of the result of the three models involved the amount of variation in the dependent variable accounted for by each model. One issue that had to be addressed before such a comparison could be made is: What value from the logistic regression analysis would be appropriate to compare to the coefficient of determination (R^2) values obtained for the linear and polynomial OLS models. Menard (2000, p. 24) indicated "[the] R^2_L [Cox-Snell R^2 value] has the most intuitively reasonable interpretation as a proportional reduction in error measure, parallel to R^2_O [coefficient of determination value used in OLS] analogs." Thus, to assess the degree of model fit we compared the Cox-Snell values produced for the logistic regression models and the R^2 values produced for the linear and polynomial OLS models.

As can be seen in Table 1, the results of the analyses are very similar. Under the high correlation condition, the goodness-of-fit values of the linear model (.701), the polynomial model (.764), and the logistic model (.660) are similar. The goodness-of-fit values were also similar for the three models under the medium correlation condition. Specifically the goodness-of-fit values for the linear, polynomial, and logistic models were .198, .199, and .196, respectively. The statistical test for each of these goodness-of-fit values was statistically significant at the .001 level, as was the case under the high correlation condition.

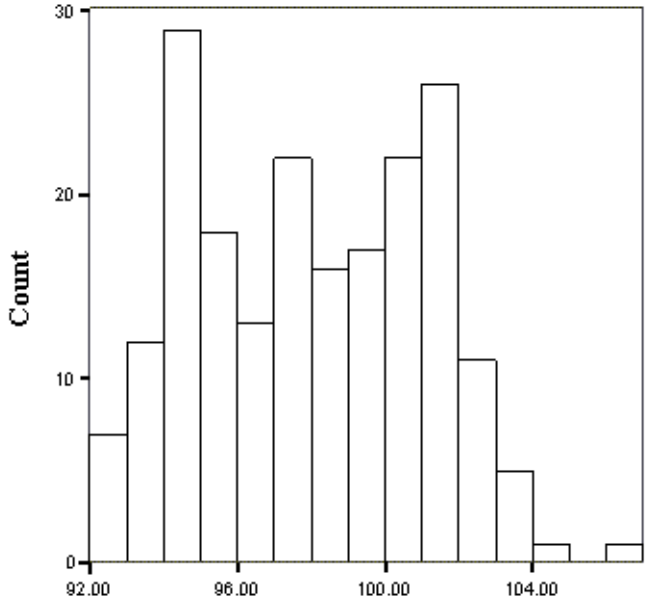

Figure 1. Bimodal distribution with high correlation (r = .837)

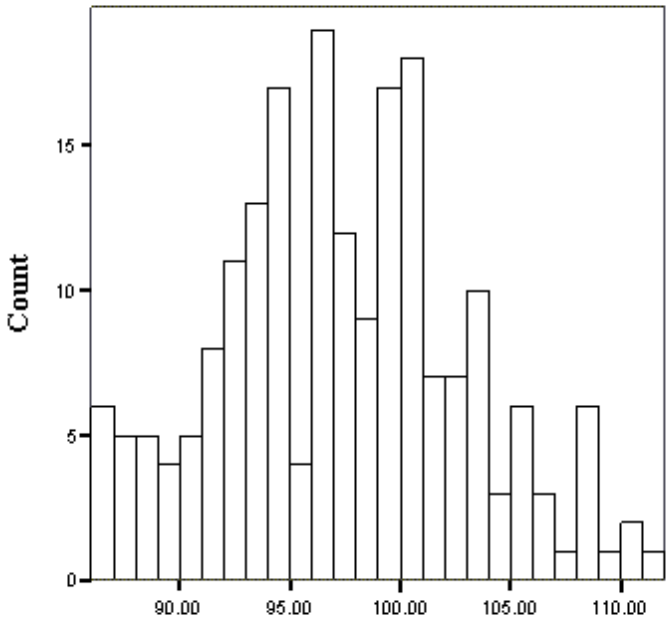

Figure 2. Normal distribution with moderate correlation (r = .445).

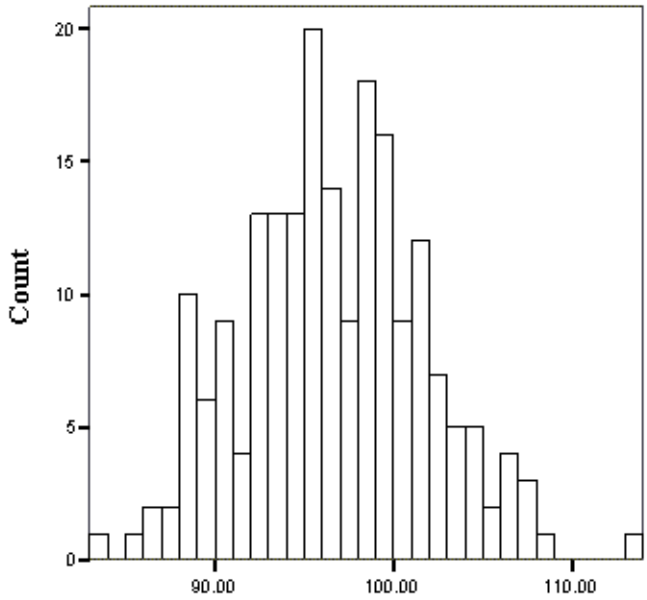

Figure 3. Normal distribution with low correlation (r = .150).

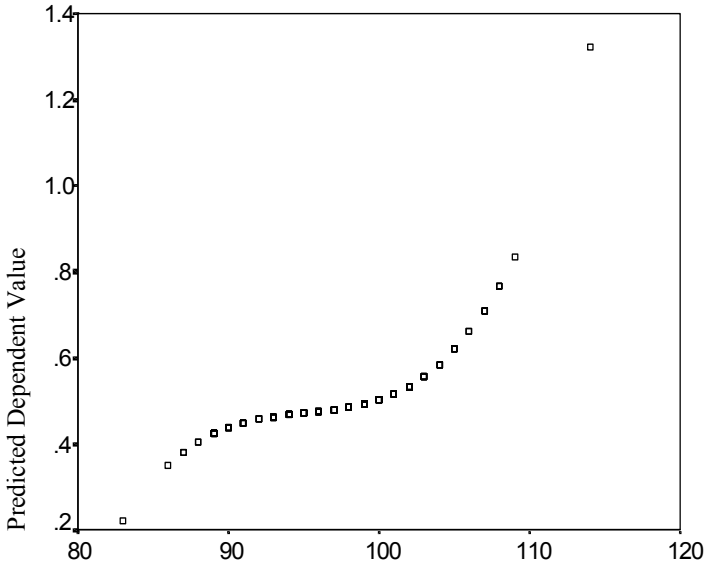

Figure 4. Predicted values of the polynomial model under the low correlation condition.

Table 1. Results of the Comparisons of the Tests of Significance Model and Condition

Bimodal Distribution (High Correlation)	*F*	*p*	R^2	Adj. R^2
Linear Model	464.96	<.001	.701	.700
Polynomial (cubic) Model	211.05	<.001	.764	.760
Logistic Model	χ^2	*p*	Cox-Snell	
	15.932	<.001	.660	
Normal Distribution (Med. Correlation)	*F*	*p*	R^2	Adj. R^2
Linear Model	48.80	<.001	.198	.194
Polynomial (cubic) Model	16.27	<.001	.199	.187
Logistic Model	χ^2	*p*	Cox-Snell	
	43.649	<.001	.196	
Normal Distribution (Low Correlation)	*F*	*p*	R^2	Adj. R^2
Linear Model	4.55	.034	.023	.018
Polynomial (cubic)	2.30	.079	.034	.019
Logistic Model	χ^2	*p*	Cox -Snell	
	4.548	.030	.022	

Under the low correlation condition similar goodness-of-fit values were obtained for the linear model (.023), polynomial model (.034), and the logistic model (.030). The statistical tests of the goodness-of-fit values for the linear and logistic models were significant at the .05 level. The statistical test of the R^2 value for the polynomial model was not significant, however, at the .05 level ($p = .079$).

Although these estimates are similar, this is perhaps the least desirable manner to compare the modeling methods as the estimates of relationship have different meanings under the OLS and logistic methods. Estimates of probability and errors in estimation of probability may be more adequate methods of comparison for these methods.

Correlations of the Predicted Probabilities

In order to compare predictions of probabilities, the dependent variable was transformed to reflect the observed probability under each of the independent variable conditions. These values formed the dependent variables used with the linear and polynomial models. If the three models were equally effective, they should produce similar estimates of probability, and the correlations between the estimates of probability under each of the conditions should be high. The correlation coefficient values for the three sets of predicted probability values are listed in Table 2. As expected, the correlation between the estimated probabilities generated by the logistic and polynomial models was higher ($r = .968$, $p < .01$) than the correlation between the estimated probabilities generated by the linear and logistic models ($r = .929$, $p < .01$).

Mean Square Error Values

Although it is clear from these correlations that there is a great deal of similarity in estimates of probability under each of these conditions, some differences emerge when one compares the mean square error values of the three models (see Table 3). As can be seen in Table 3, the models produce similar mean square error values when there is a normal distribution with a medium or low degree of relationship between the independent and dependent variables. Differences existed, however, in the mean square error terms when a larger relationship existed between the independent and dependent variables and the independent variable had a bimodal distribution.

Table 2. Correlations of Predicted Probabilities

Regression Model and Condition	Linear	Cubic	Logistic
High Correlation			
Linear	-	.941[a]	.929[a]
Cubic		-	.968[a]
Medium Correlation			
Linear	-	.988[a]	.993[a]
Cubic		-	.993[a]
Low Correlation			
Linear	-	.814[a]	1.00[a]
Cubic		-	.804[a]

[a] $p < .001$.

Table 3. Mean Square Error Values for Each Condition

Model and Condition	Sum of Squared Error	Mean Square Error
Bimodal Distribution (High Correlation)		
Linear Model	6.17	.0386
Polynomial (cubic) Model	1.82	.0091
Logistic Model	.75	.0038
Normal Distribution (Medium Correlation)		
Linear Model	4.74	.0237
Polynomial (cubic) Model	4.51	.0226
Logistic Model	4.74	.0237
Normal Distribution (Low Correlation)		
Linear Model	5.09	.0255
Polynomial (cubic) Model	4.52	.0226
Logistic Model	5.12	.0256

Table 4. Group Membership Classifications and Errors

Model and Condition	Correct Classification	False Positives	False Negatives
Bimodal Distribution (High Correlation)			
Linear Model	185	7	8
Polynomial (cubic) Model	185	7	8
Logistic Model	185	7	8
Normal Distribution (Medium Correlation)			
Linear Model	135	34	31
Polynomial (cubic) Model	135	34	31
Logistic Model	135	34	31
Normal Distribution (Low Correlation)			
Linear Model	114	39	47
Polynomial (cubic) Model	111	27	62
Logistic Model	114	39	47

Accuracy of the Classifications of Group Membership

Accuracy of the group membership classifications was the last method used to compare the three types of models. The results of the group classifications for each model under each condition are listed in Table 4. Although the predicted probabilities produced by the various models were not exactly the same, each of the models produced the exact same group membership classifications under the high and medium correlation conditions. In the low correlation condition, however, an interesting difference in the classification patterns emerged. Under this condition, the polynomial model had the lowest mean square error value but it had three more classification errors than either the linear model or the logistic model. In addition, the types of errors (false-positive errors and false-negative errors) made by the polynomial model were different from either the linear or logistic model. The polynomial model, under the low correlation condition, made the fewest false positive classifications, but it made substantially more false negative identifications than did either the linear model or the logistic model.

The differences in classification of the sample subjects could be taken as an argument against the polynomial function. This would be a valid conclusion if the sample parameters of false-positive and false-negative identifications were representative of the Type I and Type II error rates in the population as a whole. The probabilities of the subjects, who were classified differently, were located around the cut-value of .50. In the case of the polynomial model, the regression line is virtually horizontal at the mid-point and, therefore, has little predictive (discriminant) power (see Figure 4).

Discussion

The following three main issues regarding the method and results of the study need to be addressed: (a) the distributions chosen for the study, (b) the comparisons made, and (c) implications of the results regarding group membership classifications. One could argue that the distributions chosen for the present study were more alike than different. Certainly, more dramatic differences in distribution shape could have been generated. The distributions were generated with the intent of varying the degree of relationship between the independent and dependent variable (r = .837 to .150) and the shape of the independent variable distribution in terms of the number of modes (one or two) and variability around the mean (s = 3.154 to 5.636). Given the similarities in the distribution, one might have expected very similar patterns of outcomes in the comparison of the different modeling techniques. Yet, this was not the case.

An examination of the results produced by the three types of models revealed some interesting similarities and differences. Initially, we had posited that the third-degree polynomial and logistic methods would produce more comparable results because the third-degree polynomial modeling would allow the regression line to take a sigmoid shape analogous to that of the logistic model. Brown and Newman (2002) had found that polynomial modeling could, in some instances, be superior to modeling with a known (e.g., cosine) function, and it was expected this could be the case with the logistic model as well. In this study the linear, logistic, and third-degree polynomial modeling methods produced similar goodness-of-fit values. This is not, however, the only means by which the effectiveness of the models can be gauged.

Further comparisons can be made by examining the predicted probabilities and the errors in the predicted probabilities. In this regard, the third-degree polynomial and logistic models produced, as expected, more similar results than did the logistic and linear models across each of the three comparison distributions. However, the group classifications produced by these methods did not follow this pattern. Surprisingly, the logistic and linear models produced identical classifications for each of the distribution conditions despite differences in predicted probabilities. The linear, logistic and third-degree polynomial models produced identical classifications in two of the three conditions (high and medium correlations), but in the low correlation condition, the third-degree polynomial model produced three (3.5%) more errors and produced a different pattern of errors (false-positive and false-negative errors) than did the logistic and linear models.

This last comparison, the pattern of classifications, resulted in the most interesting contrast. Not surprisingly, the cases that were classified differently occurred near the cut-value of the predicted probabilities. This finding points to a concern regarding the stability of the predicted classifications for these modeling techniques. Cases that have predicted probabilities that are above the cut-value (irrespective of differences in the predicted probability) are grouped together as are cases below the cut-

value. This aggregation of cases into categories does not take into consideration the error in the predicted probability and, in turn, the stability of the model. The differences in classification produced by the models in the low correlation condition points to the need to develop a method to estimate the stability of these models when using them for classification purposes. Studies that fail to take this into consideration may unintentionally convey a greater sense of stability in the classifications provided by the models in the study. We would suggest replication as one possible means to estimate the stability of the model (Newman, McNeil, & Fraas, 2003). We strongly believe the argument between statistical significance and practical significance is not as salient as this issue of replicability; therefore, we suggest future research may wish to develop methods of estimating the stability (i.e. replicability) of the classifications produced by logistic models. To this end, the authors are presently working on a method of estimating the stability of the predicted probabilities using confidence intervals around the predicted scores and comparing the stability estimates across the three methods described in this paper.

References

Brown, R., & Newman, I. (2002). A discussion of an alternative method for modeling cyclical phenomena. *Multiple Linear Regression Viewpoints*, *28*(1), 31-35.

King, J. (April, 2002). *Logistic regression: Going beyond point-and-click*. Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA).

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*(1), 17-24.

Newman, I., McNeil, K., & Fraas, J. (April, 2003). *Deja Vu: Another Call for Replications of Research, Again*. Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL).

Peng, C., Lee, K., & Ingersoll, G. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research*, *96*(1), 3-13.

Peng, C., & So, T. (2002). Logistic regression analysis and reporting: A primer. *Understanding Statistics*, *1*(1), 31-70.

Peng, C., So, T., Stage, F., St. John, E. (2002). The use and interpretation of logistic regression in higher education journals: 1988-1999. *Journal of Research in Higher Education*, *43*(3), 259-293.

Pohlmann, J., & Leitner, D. (2004). *A comparison of ordinary least squares and logistic regression*. Manuscript submitted for publication.

Send correspondence to: Isadore Newman, Ph.D.
Department of Educational Foundations and Leadership
University of Akron
Akron, OH 44325
Email: inewman@uakron.edu

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

Mercedes K. Schneider
Ball State University

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

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

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

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

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

Purpose of the Study

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

Procedures

This Monte Carlo simulation was executed using PROC IML in SAS. Two p-dimensional, multivariate population matrices were generated, with each being $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (SAS Institute, 1999). The general procedure was as follows: In all cells, $\boldsymbol{\mu}_1$ was a p x 1 null vector, and $\boldsymbol{\mu}_2$, a p x 1 vector of effects of some combination such that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$. A sample of dimension n x p was then drawn from each population ($n_1 = n_2$; $p_1 = p_2$) and analyzed as a two-way MANOVA using Wilks' Λ and a special case of Bartlett's V as a test of significance:

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

where Λ is also calculated using a modified formula

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

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

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

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

Each n x p cell was replicated 5,000 times. For the replications where the MANOVA null hypothesis $\mathbf{H_0: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2}$ was correctly rejected within each cell, the p x 1 vectors of total and within SCs, $\mathbf{s_T}$ and $\mathbf{s_W}$, were calculated. An expected pattern across $\mathbf{s_T}$ and $\mathbf{s_W}$ was identified for each mean effect vector $\underline{\boldsymbol{\mu}_2}$ and the proportions of $\mathbf{s_T}$ and $\mathbf{s_W}$ vectors conforming to the identified pattern calculated. These proportions of $\mathbf{s_T}$ and $\mathbf{s_W}$, vectors were then compared across all levels of n, p, and $\mathbf{P}$. The use of identified SC patterns will be discussed shortly.

Three-tiered Ranking of Relative Variable Importance

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

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

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

Identified SC Patterns

The identified pattern for a given mean vector effect $\boldsymbol{\mu}_2$ was chosen by examining the *range* of SC values corresponding to each element of $\boldsymbol{\mu}_2$ when $n = 1000$ across both types of SCs and all levels of $\mathbf{P}$. A vector of least restrictive ranges for each of the p elements was then constructed to represent the identified pattern of ranges of SCs for the corresponding mean effect vector $\boldsymbol{\mu}_2$. The decision to use $n = 1000$ for identifying SC ranges is based upon the author's prior experience. When the group size is $n = 1000$ in two-group DDA, the SCs are sufficiently stabilized for interpretation of relative variable importance given the three-tiered ranking. Specifically, the ranges of SCs reflecting the effects of 0, .5, and 1 SD, respectively, are mutually exclusive. Increasing the group size n to 5000, for example, would have produced narrower ranges; however, such adjustment was not deemed necessary to term a coefficient value as indicating a ranking so general as *moderately influential*, for example.

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

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

where there are p elements. (Note: If $(k - 1) \geq p$, the proportions of *trace* of the correlation matrix, either **R** or **W**, as defined previously, accounted for by the p SC vectors would sum to one [Cooley and Lohnes, 1971]). This condition has implications for assigning unchanging SC values to indicate an effect of, say, 1 SD in the effect vector, $\boldsymbol{\mu}_2$. A variable on which the two groups differ by 1 SD will have a higher SC value in a vector of dimension p x 1 if the remaining $p - 1$ variables differ by .5 SD than if the remaining $p - 1$ variables also reflect a difference between groups of 1 SD. For example, compare the elements of the two 4 x 1 total SC ($\mathbf{s_T}$) vectors in Illustration 1 below. These two vectors have been written in the form of the condition outlined in Equation 4 above. In each, the last, bolded element is the referent kept constant at an effect of 1 SD. The remaining 3 elements represent an effect of .5 SD in the first vector and 1 SD in the second vector.

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

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

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

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

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

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

Identified SC Range:

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

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

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

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

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

Identified SC Range:

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

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

Results

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

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

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

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

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

Population Mean Vector: $\boldsymbol{\mu}_2 = \begin{bmatrix} .5 \\ 1 \end{bmatrix}$

Identified SC Range:

$$\begin{bmatrix} .34 - .54 \\ .84 - .94 \end{bmatrix}$$

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

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

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

Population Mean Vector: $\boldsymbol{\mu}_2 = \begin{bmatrix} .5 \\ .5 \end{bmatrix}$

Identified SC Range:

$$\begin{bmatrix} .58 - .81 \\ .58 - .81 \end{bmatrix}$$

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

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

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

Population Mean Vector: $\boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Identified SC Range:

$$\begin{bmatrix} .63 - .78 \\ .63 - .78 \end{bmatrix}$$

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

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

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

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

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

Identified SC Range:

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

* Vectors were either all positive or all negative

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

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

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

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

Identified SC Range:

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

* Vectors were either all positive or all negative

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

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

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

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

Identified SC Range:

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

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

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

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

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

Identified SC Range:

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

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

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

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

Population Mean Vector: $\boldsymbol{\mu}_2 = \begin{bmatrix} .5 \\ 1 \\ 1 \end{bmatrix}$

Identified SC Range:

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

* Vectors were either all positive or all negative

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

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

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

Population Mean Vector: $\boldsymbol{\mu}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$

Identified SC Range:

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

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

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

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

Population Mean Vector: $\boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ .5 \\ 1 \end{bmatrix}$

Identified SC Range:

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

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

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

One additional condition was examined where $p = 3$, that of all three variables contributing at the three different levels of effects (0, .5, and 1 SD) (Table 12). For this condition, $\mathbf{s_W}$ proportions were

consistently higher than $\mathbf{s_T}$. It is interesting to note that $\mathbf{s_W}$ proportions are consistently higher for all three conditions where at least half of the *p* variables did not contribute to group separation (i.e., the effect equals 0 SD; see Tables 1, 2, and 11).

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

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

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

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

Discussion

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

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

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

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

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

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

* Vectors were either all positive or all negative

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

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

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

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

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

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

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

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

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

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

* Vectors were either all positive or all negative

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

* Vectors were either all positive or all negative

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

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

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

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

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

* Vectors were either all positive or all negative

Issues of Practical Significance

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

Accessing a Model's Ability to Classify Subjects: The Importance of Considering Marginally Accurate Classifications

Russell Brown Cleveland State University
Isadore Newman University of Akron
John W. Fraas Ashland University

The purpose of this paper is fourfold: (a) discuss the importance of considering "marginally accurate" classifications, which are predicted probability values whose confidence limits contain the cut-value used to classifying subjects, (b) present a six-step calculation procedure used to identify the "marginally accurate" classification values, (c) illustrate how the identification of these "marginally accurate" values are important in the evaluation of the model, and (d) discuss how a review of the "marginally accurate" values can used to access the differential effectiveness of various modeling procedures with respect to their replicability and stability.

Fraas and Drushal (2004) suggested that program evaluators and educational researchers frequently encounter situations in which the dependent variable of interest is dichotomous (i.e., a variable that consists of two categories). One goal of analyzing a dichotomous dependent variable is to obtain a model that can be used to classify subjects into either of the two categories of the dependent variable. It is not uncommon for researchers and program evaluators to use a logistic regression model for this purpose. Fraas and Newman (2003) noted such classifications can be obtained from a linear probability regression model as well as a logistic regression model. In addition, Brown, Newman, and Fraas (2004), explain how a third-degree polynomial regression could be used to classify each subject into one of the two categories of the dependent variable.

Regardless of which analytic method is used, program evaluators and researchers need to consider an issue that is often overlooked. That is, how confident are they in the classification of the subjects? The purpose of this paper is fourfold: (a) discuss the importance of considering "marginally accurate" classifications, which are predicted probability values whose confidence limits contain the cut-value used to classifying subjects, (b) present a six-step calculation procedure used to identify the "marginally accurate" classification values, (c) illustrate how the identification of these "marginally accurate" values are important in the evaluation of the model, and (d) discuss how a review of the "marginally accurate" values can used to access the differential effectiveness of various modeling procedures with respect to their replicability and stability.

Need for Additional Classification Table Information

The need for providing information that can be used to supplement the classification table produced by analytic methods used in conjunction with a dichotomous dependent variable occurred to us when we attempted to compare the results of the three analytic techniques (Newman, Brown, & Fraas, 2004). We found that misclassifications became problematic in trying to explain the comparative results of three methods. Although similar results were obtained by the three different models when comparing the methods in terms of tests of significance and predicted probabilities, some differences existed in the group classifications produced by them. Under a condition in which there was a modest correlation between the independent variable, the third-degree polynomial model produced 3.5% more errors than did the logistic or linear models, which produced identical classification patterns. A closer examination of these differences in classification showed the cases that were classified differently had predicted probabilities that were all close to the cut-line probability level of .50.

We believe the degree of confidence we have in two models that assign the same classification to each subject may or may not be the same. If the first model produces predicted probabilities for the subjects that have greater variation than those of a second model and a number of those probabilities are located near the cut-line of .50, our confidence in the first model will not be as strong as it is in the second model even though both models classified the subjects the same.

To address this issue, we believe information conveyed by the classification table, which lists the number of subjects correctly classified and incorrectly classified, should be supplemented by reporting the number and percentage of classifications that are "marginally accurate." If the number or percentage of such classifications for a model is small, program evaluators and researchers would have greater confidence in using the model to classify future subjects. The discussion presented in the next section provides the steps researchers need to complete in order to access a model in such a fashion.

Method

A method is presented through an illustration that can be used to identify the number and percent of "marginally accurate" values (i.e., predicted student probability values whose confidence limits contain the cut-value used to classifying subjects). The illustration used is taken from Fraas and Drushal (2004) in their discussion of the use of delta-p values as a means to understand the effect of incremental changes in the independent variables on the predicted probability in a logistic model.

The data from the Fraas and Drushal (2004) study contained information on 525 college students. They were "interested in assessing the relationship between various student and financial factors recorded for students who have applied to a university and whether the students actually did or did not matriculate" (p. 5). The dependent variable indicated in which of two categories each student belonged. Each student who did not matriculate was assigned a value of one, while each student who did matriculate was assigned a value of zero. The independent variables used to predict whether or not an individual student did or did not matriculate were as follows:

1. The students' high school grade point averages (HSGPA)
2. The students' ACT composite scores (ACT)
3. The sex of each student (SEX) [0 = female student; 1 = male student]
4. The amount of financial aid offered each student (AID)
5. The amount of financial need established for each student (NEED).

A linear probability model, a third-degree polynomial model, and a logistic regression model were used to analyze the relationships between the independent variables and the dependent variable. In order to be able to produce the terms for the third-degree polynomial model, a multiple linear regression was used to produce a single standardized weighted predicted composite score for each subject. This score was then squared and cubed to produce the terms for the third-degree polynomial model.

Each regression method was subsequently used to establish predicted probabilities and predicted classifications for each of the subjects in terms of the dependent variable (i.e., was the subject predicted to matriculate or not matriculate). Subjects whose predicted probability values were greater than or equal to .5 were classified as having matriculated; while subjects whose predicted probability values were less than .5 were classified as not having matriculated. The classification results of these analyses can be seen in Table 1.

As one can see from the results listed in Table 1, the percent correctly classified by the three methods are quite similar. The polynomial model produced the greatest number of correct classifications (58.1%). The linear and logistic models produced an equal percentage of correct classifications (57.3%), but produced a different pattern of false positive and false negative identifications.

The specific issue we are attempting to address is: What number and percent of the classifications are "marginally accurate" (i.e., "unstable")? The calculation of the number and percent of correctly classified subjects whose classifications are "marginally accurate" can be calculated in six steps regardless of which model is used.

The calculations used in our illustration are for the logistic regression model. The required steps are as follows:

1. The two standard deviation values for the predicted probability values--one for the group of students who were classified by the model as not matriculating and the other for the group of students who were classified by the model as matriculating--are calculated. The standard deviation values for the groups of students who did not and did matriculate were .053 and .048, respectively.

2. Since we are interested in a one-tailed limit value for each group, the standard deviation value for each group is multiplied by 1.65 (the t value for the one-tailed 95% confidence level). Thus the value for the students who were classified by the model as not matriculating was .088 (.053 X 1.65), while the value for the students who were classified by the model as matriculating was .079 (.048 X 1.65).

Table 1. Original Group Membership Classifications and Errors

Model	Correct Classification		False Positives	False Negatives	Percent Correct
	1	0			
Linear Model	194	107	145	79	57.3%
Polynomial Model	185	120	132	88	58.1%
Logistic Model	195	106	146	78	57.3%

3. The value of .088 was added to each predicted probability for the students classified by the model as not matriculating; while .079 was subtracted to each predicted probability for the students classified by the model as matriculating.

4. The number of students who were classified by the model as not matriculating but whose upper predicted probability limit values equaled or exceeded .50 was recorded. A total of 35 students had upper limits that equaled or exceeded .50. Thus of the original 106 who were correctly classified as not matriculating, 33.0% had upper predicted probability limits that equaled or exceeded .50. We labeled these classifications as "marginally accurate."

5. The number of students who were classified by the model as matriculating but whose lower predicted probability limit values fell below .50 was recorded. A total of 71 students had lower limits that fell below .50. Thus of the original 195 who were correctly classified as not matriculating, 36.4% had lower predicted probability limits that fell below .50. Again, we labeled these classifications as "marginally accurate."

6. The total number and total percent of correctly classified students who were labeled as "marginally accurate" were noted. The total number was 106 (35 + 71) and the total percent was 35.2% [((35 + 71)/ 301) X 100].

The number and percent of students labeled "marginally accurate" or "unstable" were calculated in the same manner for the linear probability and the third-degree polynomial models. See Table 2 for the results of those calculations.

Regardless of which model is used, we suggest that the percent of "marginally accurate" figures (e.g., 33.0% of the students correctly classified as not matriculating; 36.4% of the students correctly classified as matriculating; and 35.2% of the students overall correctly classified for the logistic regression model) should be reported along with the classification table that is normal provided by an analysis of a dichotomized dependent variable.

An examination of the number of subjects "marginally accurate" for each of the three types of models may lead researchers to reach a different conclusion regarding the desirability of using a given model than would a review of the number of subjects correctly classified by each model is reviewed. The number of subjects correctly classified by each model (see Table 1) would suggest that the models are approximately equally effect in classifying students. A review of the number of subjects identified as "marginally accurate" would indicate that the polynomial model, which had the lowest number of "marginally classified" subjects (see Table 2) may be the preferred model. Thus it may be important, both in a relative and an absolute sense, for researchers to access both criteria (i.e., the number and percent "marginally accurate" as well as the number and percent correctly classified) when accessing a models ability to classify students.

Discussion

When attempting to evaluate the effectiveness of a model designed to classify subjects into one of two groups with a linear probability model, a third-degree polynomial model, or a logistic regression model researchers may find the information provided by the classification table insufficient. The application of the technique for identifying the number and percent of "marginally accurate" classifications, as presented in this paper, may be used to supplement the information presented in the standard classification table. The fewer the number of identified "marginally accurate" classifications the more confident the researchers will be in their model's ability to classify future subjects.

Table 2. Changes in Group Membership Classifications for Students Correctly Classified by the Model

Model and Original Correct Classification	Marginally Accurate	Stable	Total	% Marginally Accurate
Linear Model				
Matriculated	75	119	194	38.7%
Did not Matriculate	44	63	107	41.1%
Total	119	182	301	39.5%
Polynomial Model				
Matriculated	66	119	185	35.6%
Did not Matriculate	23	97	120	19.2%
Total	89	216	305	29.2%
Logistic Model				
Matriculated	71	125	195	36.4%
Did not Matriculate	35	70	106	33.0%

We believe that researchers often judge a model's ability to predict group membership or an occurrence of an event primarily through the use of the classification table. The values reported in the standard classification table, however, do not take into consideration the number of probability values (i.e., the values on which the classifications are based) that are close to the cut-value used to classify the subjects. If a large number of these probability values are located near the cut-value, researchers may find the accuracy of the classifications of future subjects unacceptable. That is, the model did not provide sufficient stability from sample to sample.

Researchers may find it important to identify in which classification most of the "marginally accurate" values are located (i.e., in the classification assigned the value of 0 versus the classification assigned the value of 1). If the marginal values are predominately in the classification assigned the value of 1 (the event did occur) and few or no marginal values are located in the classification assigned the value of 0 (the event did not occur), the researchers may be more confident in their classifications of the event not occurring than not occurring.

We realize there are other methods that researchers can use to access a model's ability to classify subjects (e.g., the use of a holdout group). The key issue, however, is that regardless of how researchers evaluate a model's ability to classify subjects, consideration should be given to the confidence they have in their classifications. We believe is an analytic technique that will improve data-based decision making by forcing researchers to reflect on how their models will be used and the degree of confidence they have in the use of those models. If a model is to be used to classify future subjects, the concept of "marginally accurate" values may be a key concept for researchers to consider.

References

Brown, R., & Newman, I. (2002). A discussion of an alternative method for modeling cyclical phenomena. *Multiple Linear Regression Viewpoints*, *28*(1), 31-35.

Fraas, J. W. & Newman, I. (February, 2003). *Ordinary lest squares regression, discriminant analysis, and logistic regression: Questions researchers and practitioners should address when selecting an analytic technique*. Paper presented at the Eastern Educational Research Association, Hilton Head, SC.

Fraas, J. W., & Drushal, J. (2004). *Expressing logistic regression coefficients as delta-p values*. Manuscript presented for publication.

Newman, I., Brown, R., & Fraas, J. W. (April, 2004). *Logistic regression as compared to linear and polynomial least squares regression: Is OLS 3rd degree polynomial a more fair comparison to the logistic method*. Paper presented at the American Educational Research Association, San Diego, CA.

Send correspondence to: Russell Brown, Ph.D.
The University of Akron
The Polsky Building 315f
225 S. Main Street
Akron, OH 44325-4207
Email: rcbrown@uakron.edu

Bootstrapping within the Multilevel/Hierarchical Linear Modeling Framework: A Primer for Use with SAS and SPLUS

J. Kyle Roberts
University of North Texas

Xitao Fan
University of Virginia

Nested data structure obtained from a cluster sampling design often calls for hierarchical linear modeling (HLM) analysis. Such data structure warrants some special considerations when the bootstrap technique is applied. This paper presents some discussions and examples for applying the bootstrap method within the framework of hierarchical linear modeling. A two-level dataset (about 900 students nested under 20 schools) extracted from the High School and Beyond (HSB) was used for illustration. Bootstrap resampling was implemented in both SAS and S-PLUS, and a hierarchical linear model with one Level-1 predictor (student SES), and one Level-2 predictor (type of schools, Catholic or public) was applied.

In quantitative research in education and psychology, over-reliance on statistical significance testing has been called into question. Several issues have been raised concerning the use of statistical significance testing in research practice including sample size, the meaningfulness of the traditional null hypothesis, and questions involving the validity of theoretical assumptions underlying parametric statistical inferences (e.g., Carver, 1978; Shaver, 1993; Thompson, 1993). As a result of these and other concerns, researchers are increasingly turning to empirically-grounded resampling procedures in quantitative analyses.

Applauded as one of the newest breakthroughs in statistics (Kotz & Johnson, 1992), the bootstrap is often considered the best-known resampling method. The importance of bootstrapping as a versatile analytic approach with which to conduct data analysis has been widely recognized not only by those in the area of statistics, but also by quantitative researchers in education, psychology, and social and behavioral sciences in general.

Statistical inference (e.g., in a t test, rejection of the null hypothesis that two populations have equal means) is usually made based on the sampling distributions of a statistical estimator. For parametric statistics, the derivation of such sampling distributions is typically based on a set of theoretical assumptions. The bootstrap method attempts to estimate these sampling distributions empirically, using information drawn from the sample of observations used to estimate the statistical model in the first place (Diaconis & Efron, 1983; Efron, 1979). In doing so, the bootstrap approach avoids some of the pitfalls of traditional statistical significance testing. As discussed by Lunneborg (2000):

> Until inexpensive computing power made replicate data analysis practical, the drawing of statistical inferences from a set of data almost always required that we accept an idealized model for the origin of those data. Such models can be either inappropriate or inadequate for the data in our study. Resampling techniques allow us to base the analysis of a study solely on the design of that study, rather than on a poorly-fitting model. (p. xi)

The bootstrap method has found a variety of research applications in social and behavioral sciences. For example, the bootstrap method has been applied in sociological research (e.g., Stine, 1989), and in research for psychological measurement issues such as differential test predictive validity (e.g., Fan & Mathews, 1994) and item bias (e.g., Harris & Kolen, 1989). Application of bootstrapping has also involved many different statistical techniques, including correlation analysis (e.g., Mendoza, Hart, & Powell, 1991; Rasmussen, 1987), regression analysis (e.g., Fan & Jacoby, 1995), descriptive discriminant analysis (e.g., Dalgleish, 1994; Thompson, 1992), canonical correlation analysis (e.g., Fan & Wang, 1996; Thompson, 1995), factor analysis (e.g., Lambert, Wildt, & Durand, 1991; Thompson, 1988), and structural equation modeling (e.g., Bollen & Stine, 1990; Yung & Bentler, 1996).

Although bootstrap was proposed as a versatile tool for non-parametric statistical inference (Efron, 1985), Thompson (1993) has also advocated the use of bootstrapping as a descriptive tool and an internal replication mechanism for assessing the stability and replicability of sample results of an individual study. This descriptive use of bootstrap is meaningful when our interest is not about statistical inference, but rather, about understanding how stable the results may be across repeated sampling.

Bootstrapping is a computing-intensive data resampling strategy, and easy access to powerful computing facilities makes bootstrapping an attractive and viable procedure for research practitioners. Unfortunately, although the logic of bootstrapping is conceptually straightforward, bootstrapping has yet

to enjoy widespread application in many areas of research and for some statistical techniques. Because bootstrapping is not typically implemented as an automated option in the major commercial statistical software packages (e.g., SAS, SPSS), researchers who desire to use this approach usually have to deal with programming for performing bootstrap resampling. This can be a daunting endeavor for many who do not have the skills, knowledge, or interest required to carry out such a task. Consequently, this appears to be a major obstacle for implementing bootstrapping in substantive research. Some methodologists have sensed the need for programs to perform bootstrapping; as a result, some special programs have been published for bootstrap application in different analytic techniques, such as regression analysis (Fan & Jacoby, 1995) and factor analysis (Thompson, 1988). But overall, bootstrapping remains procedurally difficult for most research practitioners.

Multilevel modeling is an area where bootstrapping has not yet enjoyed much application. As is the case for other statistical techniques, bootstrapping within multilevel modeling may serve two main purposes: making non-parametric inferences about parameter estimates and correcting potential bias in parameter estimation. This non-parametric approach can be especially helpful in samples where assumptions about data may have been violated (e.g., data non-normality, Bryk & Raudenbush, 1992), or in samples where the number of Level 1 observations (e.g., individuals) may be small within each Level 2 unit (e.g., schools).

This paper provides some heuristic examples of implementing bootstrap analysis for hierarchical linear modeling (HLM). Although there has been little application of bootstrapping in hierarchical linear modeling, it is hoped that the demonstration of the use of these methods will encourage future researchers to utilize these techniques. Procedures for conducting bootstrapping analyses with both SAS and S-PLUS are presented with heuristic datasets.

Bootstrap Approach

Bootstrap as the most popular resampling method is mainly used for estimating the sampling distribution of a statistic of interest for which parametric alternatives either do not exist, or the validity of the parametric alternatives are in question (e.g., violated assumptions). The basic bootstrap method typically has three straightforward steps:

1. select *B* independent bootstrap samples, each consisting of n observations drawn with replacement from the original sample,
2. obtain the statistic of interest from each bootstrap sample, and
3. evaluate the sampling distribution of the bootstrapped statistic of interest by
    a) estimating the standard error of the statistic of interest by the sample standard deviation of the *B* bootstrap replications, or
    b) using exact percentiles (e.g., 97.5%; 2.5%) for constructing empirical confidence intervals.

Approach a) above assumes distribution normality of the bootstrapped statistic, and parametric confidence intervals can be constructed through the use of the estimated standard error. Approach b), however, does not assume distribution normality of the bootstrapped statistic, and the resultant confidence intervals are non-parametric in nature.

Although the bootstrapping method as described above is procedurally straightforward, its application in hierarchically nested data structure such as those used in hierarchical linear modeling may warrant some special considerations. Typical bootstrapping involves sampling individual observations with replacement, and there is no consideration for the nested data structure in HLM, (e.g., individual students (Level 1 units) are nested under schools (Level 2 units)). Because of this nested data structure, potentially, there can be different resampling approaches for hierarchically nested data. From a sample data with two levels (e.g., Level 1: students, and Level 2: schools), with k schools, and each with ni students, and the total sample size of N [$N = \mathbf{\Sigma} n_i$, i = 1, 2, 3 . . j, k], the following bootstrap sampling approaches may potentially be applied:

1. draw a bootstrap sample of *N* students with replacement, and totally ignore the nested data structure;
2. draw a bootstrap sample of ni students with replacement from each and every school in the sample data, and the bootstrap sample has sample size of *N*;
3. bootstrap *k* schools with replacement while selecting all n_i students in each bootstrapped *k* school;
4. first, drawn a bootstrap sample of *k* schools with replacement; from each sampled school, draw a bootstrap sample of n_i students with replacement.

The first two approaches will provide a consistent sample size of *N* for each bootstrap iteration. But the third and fourth approach will not provide a consistent sample size of *N* for each bootstrap iteration, unless $n_i = n$ for each Level 2 unit (i.e., each school contains the same number of students in the original sample).

Theoretically, both Level 2 and Level 1 units should be considered as randomly drawn from the population. In other words, in clustered sampling design, Level 2 units (schools) are randomly drawn first. Level 1 units (students) are then randomly drawn from the school. In this sense, the fourth approach of bootstrap sampling for hierarchically nested data described above makes good sense. In practice, however, the fourth approach will typically not provide a consistent bootstrap sample size of *N*, because hierarchically nested sample data typically do not have equal sample size within each Level 2 unit. Without a consistent sample size of *N*, it would not be possible to construct an empirical sampling distribution for a statistical estimator of interest because the sampling distribution is always associated with a specific sample size. For this reason, we only used the first two bootstrap sampling approaches in our examples.

Data Source

Bryk and Raudenbush (1992) used a dataset from the national High School and Beyond (HSB) database to illustrate the application of HLM. The same dataset is also used by Singer (1998) in her illustration of using SAS for fitting HLM models. For bootstrapping illustrations in this paper, we used a dataset of 20 schools randomly selected from the dataset of 160 schools as used in Bryk and Raudenbush (1992) and Singer (1998).

Table 1 presents the basic descriptive information for the variables used in our HLM bootstrapping example. The student level predictor SES is centered with mean of zero. The variable SECTOR is dummy coded, with Catholic schools coded as 1 and public schools coded as 0. So the mean of SECTOR (0.35) indicates that, of the 20 schools in this dataset, seven (35%) are Catholic schools, and the remaining 13 are public schools.

Table 2 presents the descriptive information for math achievement scores for the 20 schools, and the sample sizes of the 20 schools. The total sample size for this data is 914, with sample size for individual schools ranging from 25 to 66, and the average sample size across the 20 schools of 45.7. It is noticed that there appears to be some noticeable variation among the school averages of math achievement score. This suggests that some school-level variable may potentially be useful in accounting for the variation among the school means of the math achievement score.

A conditional two-level model, with SES as the Level 1 (student level) predictor, and SECTOR as the Level 2 (school level) predictor, was fitted to the data, as shown below (Y: math achievement score; notations as used in Bryk and Raudenbash, 1992):

$$\text{Level 1:} \quad Y_{ij} = \beta_{0j} + \beta_{1j}\,(\text{SES}) + r_{ij}\,, \quad \text{and}$$

$$\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{SECTOR}) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}\,(\text{SECTOR}) + u_{1j}\,.$$

To provide information about how much variation in the math achievement score is within and between schools in this dataset, a one-way ANOVA model with random effects was fitted the data. The one-way ANOVA model with random effect takes the following form:

$$\text{Level 1:} \quad Y_{ij} = \beta_{0j} + r_{ij}\,, \quad \text{and}$$

$$\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + u_{0j}$$

Table 1. Descriptive Statistics for the Dataset Used

Variable		
Student-Level (1)	Mean	SD
Math Achievement (Y)	13.00	7.20
SES	0.00	0.65
School Level (2)		
Sector	0.35	0.49

Table 2. Descriptive Statistics of Math Achievement Scores for the 20 Schools

School ID	*N*	Mean	SD
1317	48	13.18	5.46
1374	28	9.73	8.36
1461	33	16.84	6.95
1477	62	14.23	7.15
2458	57	13.99	5.85
2629	57	14.91	5.17
2768	25	10.89	7.29
2771	55	11.84	6.80
3427	49	19.72	3.54
3716	41	10.37	8.48
3838	54	16.06	5.10
3967	52	12.04	6.89
4383	25	11.47	7.45
5619	66	15.42	7.28
5762	37	4.32	4.99
6291	35	10.11	6.59
6897	49	15.10	6.65
7697	32	15.72	6.62
7890	51	8.34	6.25
8946	58	10.38	6.52

Table 3. Results of One-Way ANOVA Model and the HLM Model

One-Way ANOVA			
Fixed Effects	Coefficients		Description
γ_{00}	12.76		overall mean math score
Random Effect	Variance Component		
School Mean, u_0	11.09		variation of school means
Level-1 Residual	41.86		
Intra-Class Correlation $\rho = 11.09/(11.09 + 41.86) = .21$			
HLM Model: (Level 1 Predictor: SES; Level 2 Predictor: SECTOR)			
	Coefficients		
Fixed Effects	SAS	S-PLUS	Description
γ_{00}	11.33	11.33	mean math score for public schools
γ_{01}	4.02	4.02	SECTOR main effect
γ_{10}	3.35	3.34	SES main effect
γ_{11}	-1.77	-1.77	**a** SECTOR effect on SES slope
	Variance Component		
Random Effects	SAS	S-PLUS	
School Mean, u_0	7.64	6.78	variation of intercept (τ_{00})
SES-Math Slope, u_1	2.54	2.07	variation of slope (τ_{11})
Covariance (u_0,u_1)	2.16	1.91	covariation of u_0 and u_1 (τ_{01})
Level-1 Residual	37.72	37.72	Var (r_{ij})

a In Catholic schools (coded as 1 on SECTOR), the student performance on Math is less related to SES (Catholic schools are more equitable). See Chapter 4 in Bryk and Raudenbush (1992) for discussion related to this issue.

Table 3 presents the results of fitting the two different models to this dataset. The first one is an unconditional model, or the one-way ANOVA model with random effect, and the second one is the HLM model we used for later bootstrapping illustration. From the one-way ANOVA model with random effect, the intraclass correlation was obtained to be 0.21, suggesting that 21% of the variance in the math scores is between-school variation, while the remaining 79% variation is within schools. This indicates that the HLM model is warranted for this dataset. If it turned out that only a negligible proportion of the total variance is between-school variation, HLM would not be as useful. Since the nested bootstrap will be illustrated with two software packages, SAS and S-PLUS, results from each of these packages will be presented in the HLM model in the following tables.

Further comparisons between the two models show that a) for the between-school variation, 31% of the variance [(11.09 – 7.64)/11.09] is accounted for by the school-level predictor SECTOR, and b) 10% of within-school variation is accounted for by the student-level predictor SES [(41.86 – 37.72)/41.86]. The results of the HLM model indicate that, within Catholic schools, the relationship between math achievement and SES is weaker than that within public schools (γ_{11} = -1.77). More specifically, for

public schools, the average regression slope between math score and SES is 3.35. For Catholic schools, the average regression slope between math score and SES is 1.58 (3.35-1.77), suggesting that Catholic schools appeared to be more equitable with regard to student SES level (Bryk & Raudenbush, 1992, Chapter 4).

Method

As has been noted previously, performing a nested bootstrap within the HLM framework is not just a "point and click" procedure in any software package. Although some programs, such as MLwiN, provide a method of performing the bootstrap with hierarchically structured data, this method is based on residuals bootstrap, which redistributes the residuals at each appropriate level (see bootstrap #4 above) rather than nesting the bootstrap within Level 2 units.

The nested bootstrap utilizes a nested looping structure within both the SAS and S-PLUS architecture. Inside the inner loop, a dataset is being created from the original dataset by extracting the data, one school (or Level 2 unit) at a time. The programs will search through the data and find the first appearing school and then extract all other pieces of data that have the same value for the school variable. In the case of the HSB dataset, 20 total schools were selected. The dataset, after being split into the 20 schools, is bootstrapped across the data contained in each school such that the number of people in the original school equals the number of people in the now bootstrapped school. The iterative process can be described as follows:

1. Select all data in school k.
2. Bootstrap data in school k such that $n_i = n_i'$.
3. Repeat process for next k school.
4. Append data from school $k + 1$ to school k.
5. Repeat steps 3 and 4 until all schools have been selected.

After this inner loop has created the bootstrapped dataset, the HLM analysis is conducted in an outer loop and the desired components are extracted. This outer loop then reverts back to the original dataset and the entire process is begun again. In the outer loop, the extracted components are appended to the previously extracted components across all bootstrapped samples. In the case of this paper, we chose 2000 bootstrap samples. It should be noted that these two programs are computer intensive and require between 2 and 3 hours of processing time for 2000 iterations on a Pentium III 600 MHz with 128meg RAM.

For comparison purposes, we also chose to include in the analysis a typical (non-nested) bootstrap. In this method, student scores were bootstrapped regardless of which school they appeared in (see bootstrap method #1 above). It is conceivable that in this method, within a single bootstrap, one school may contain no student estimates for a given bootstrap sample. This analysis was only conducted in SAS and as such, should be compared against the original SAS estimates.

Results

Criteria were set for which pieces of information that should be extracted from the HLM analysis from the specified model. Since this was a model with two levels and random effects at the second level, four fixed effects, two random effects, the covariance between the random effects, and the Level 1 residual were extracted in each bootstrapped sample. The results of these bootstrapped fixed and random effects can be seen in Tables 4 and 5.

In Tables 4 and 5, columns labeled "Original Data SAS" and "Original Data S-PLUS" correspond to the results from the original sample analysis in Table 3. These were included for comparison purposes. Results from the SAS nested bootstrap program, the S-PLUS nested bootstrap program, and from the non-nested bootstrap are also included in these tables.

In first looking at the results from Table 4, it can be seen that the γ_{00} and γ_{01} fixed effects had bootstrapped sample estimates that differ only slightly from the original estimate. This was not the case, however, with the bootstrapped estimates for g10 in the S-PLUS bootstrapped estimate and for γ_{11} in both the SAS and S-PLUS bootstrapped estimate. In these later estimates, it can be seen that the estimate 95%

Table 4 Results of the HLM Bootstrap of the HSB Data for the Fixed Effects

γ_{00} (mean math score for public schools)					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	11.33	11.33	11.33	11.33	11.33
Minimum		10.11		10.23	10.37
Maximum		12.24		12.29	12.23
SD		0.29		0.29	0.29
SEM		0.01		0.01	0.01
LCL Mean		11.32		11.31	11.32
UCL Mean		11.34		11.34	11.35
Skewness		-0.03		-0.04	-0.12
Kurtosis		0.04		0.09	-0.10
γ_{01} (SECTOR main effect)					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	4.02	4.01	4.02	4.03	4.01
Minimum		2.66		2.79	2.78
Maximum		5.25		5.44	5.34
SD		0.40		0.40	0.41
SEM		0.01		0.01	0.01
LCL Mean		3.99		4.01	4.00
UCL Mean		4.02		4.04	4.03
Skewness		0.03		0.08	0.03
Kurtosis		-0.14		0.01	-0.12
γ_{10} (CSES main effect)					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	3.35	3.36	3.34	3.37	3.37
Minimum		2.08		2.08	1.78
Maximum		4.80		4.91	4.72
SD		0.42		0.42	0.42
SEM		0.01		0.01	0.01
LCL Mean		3.34		3.35	3.35
UCL Mean		3.38		3.39	3.39
Skewness		0.03		0.06	0.01
Kurtosis		-0.07		0.04	-0.09
γ11 (SECTOR effect on CSES slope)					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	-1.77	-1.80	-1.77	-1.83	-1.81
Minimum		-3.75		-3.70	-3.72
Maximum		0.55		-0.01	-0.03
SD		0.58		0.59	0.60
SEM		0.01		0.01	0.03
LCL Mean		-1.83		-1.85	-1.84
UCL Mean		-1.78		-1.80	-1.78
Skewness		0.01		-0.02	-0.06
Kurtosis		0.08		-0.16	-0.08

Table 5 Results of the HLM Bootstrap of the HSB Data for the Random Effects

Variation of Regression Intercept u_0 (τ_{00})					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	7.64	8.61	6.78	7.61	8.55
Minimum		4.56		3.82	5.43
Maximum		13.30		12.66	12.68
SD		1.26		1.12	1.23
SEM		0.03		0.25	0.03
LCL Mean		8.56		7.56	8.50
UCL Mean		8.67		7.66	8.61
Skewness		0.14		0.25	0.22
Kurtosis		0.12		0.21	-0.11
Variation of Regression Slope u_1 (τ_{11})					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	2.54	4.43	2.07	3.68	4.43
Minimum		0.43		0.01	0.72
Maximum		11.67		8.79	11.69
SD		1.56		1.36	1.56
SEM		0.04		0.03	0.04
LCL Mean		4.36		3.63	4.37
UCL Mean		4.50		3.74	4.50
Skewness		0.56		0.37	0.45
Kurtosis		0.66		0.11	0.27
Covariation Between u_0 and u_1 (τ_{00})					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	2.16	2.33	1.91	2.07	2.34
Minimum		-1.74		-1.75	-1.37
Maximum		6.38		5.61	6.83
SD		1.15		1.02	1.11
SEM		0.03		0.02	0.03
LCL Mean		2.27		2.02	2.29
UCL Mean		2.38		2.11	2.40
Skewness		-0.05		0.02	-0.00
Kurtosis		0.17		-0.01	0.01
Level-1 Residual (Var (r_{ij}))					
Estimate	Original Data SAS	SAS Nest Boot	Original Data S-PLUS	S-PLUS Nest Boot	Bootstrap All
Mean	37.72	36.18	37.72	36.09	36.16
Minimum		31.22		30.70	30.56
Maximum		41.78		42.37	40.87
SD		1.53		1.53	1.55
SEM		0.03		0.03	0.03
LCL Mean		36.11		36.02	36.10
UCL Mean		36.24		36.16	36.23
Skewness		0.05		-0.02	0.06
Kurtosis		-0.01		0.13	-0.11

confidence intervals did not capture the original data estimates. In instances of the fixed effects estimates, it would then be proper to default to the bootstrapped estimate rather than assume the original estimate.

In Table 5 we can see the differences between the original estimates of the variance components and the bootstrapped estimates. Upon looking at these results, we can see that in every bootstrapped estimate across both the SAS and S-PLUS results, the original estimate of the variance component is not captured by the 95% confidence intervals around the nested bootstrapped estimate. This is especially troubling since on many occasions, model specification issues are often based on these estimates.

For example, consider the intraclass correlation from the original dataset and nested bootstrap (where $ICC = \tau_{00}/[\tau_{00} + r_{ij}]$). In the case of the present dataset, the ICC for the original data in the SAS equation would be .168, suggesting that only 16.8% of the variance in the math scores is between-school variation, while the remaining 83.2% variation is within schools. We can contrast these results to the nested bootstrap sample where the ICC is .192. This is critical when we consider that Kreft and de Leeuw (1998) and Roberts (2004) have defined an ICC of .20 or greater as a large effect. These differences in the variance estimates might lead a researcher to interpret a fixed effect as a small or medium effect when in fact the effect is quite large (or vice versa). This could prove problematic when basing modeling decisions on the interpretation of variance estimates alone.

Discussion

One question that might be brought to attention from the results presented in Tables 4 and 5 is whether or not the effort justifies the ends. In this analysis, the results from the individual bootstrap and the nested bootstrap yield similar results. Although this has proven true in this case, it does not hold that the two types of resampling will yield similar results across all hierarchical datasets. Consider when a dataset (unlike the present dataset) has few Level 1 units inside each Level 2 unit. In this case, the individual bootstrap would be much more likely to obtain bootstrap estimates in which entire Level 2 units are ignored, whereas the nested bootstrap will always include the same n for each Level 2 unit in every analysis. As was previously noted, the nested bootstrap will prove especially useful when *N* for the entire dataset is very small.

While the present paper has only identified certain components of the hierarchical linear model to bootstrap, it can be seen that bootstrapping other components of the model could help to answer some of the problems associated with assumptions in HLM. For example, we might wish to test the mutual independence of all residuals by testing to see if they are normally distributed and have zero means given the explanatory variables across bootstrap samples. Furthermore, we might also want to test each bootstrapped sample for heteroscedasticity and in cases where heteroscedasticity is high across bootstrap samples, apply a Box-Cox transformation to the dataset and then reapply the bootstrap (Snijders & Bosker, 1999).

Although it has been the primary purpose of this paper to discuss and illustrate the nested bootstrap in hierarchical linear and multilevel modeling, further applications of this type of analysis could be utilized beyond the topics presented currently. For example, this type of nested bootstrap could prove vitally useful in Generalizability theory studies where actual items are bootstrapped rather than just individuals. This nested bootstrap might further be utilized in ANOVA type studies where researchers are concerned about the robustness of variance estimates within levels of a given way.

This type of resampling can also encourage researchers to think seriously about resampling designs in other types of analysis. For example, consider if we were to apply a jackknife resampling design to the present study. Since HLM type analyses require such large sample sizes, we are unlikely to see much of a difference in our parameter estimates. Consider, however, if we were to apply a nested jackknife to the data where actual schools are jackknifed rather than individuals. In this case, a researcher could easily note the potential contribution (or lack of contribution) for each school in the dataset. This type of analysis could also be applied to Generalizability theory where items (or some other facet) are jackknifed rather than individuals.

This type of resampling could be further applied in a nested jackstrap (a combination of the nested bootstrap and nested jackknife). In this type of analysis (in a school-effects model), schools would first be jackknifed and then the nested bootstrap would be applied to each jackknifed sample. One might consider that this type of analysis could conceivably run on a single computer for a couple of days, but

the results could help solve some sampling issues that a researcher might be facing. It is hoped that the presentation of this paper will encourage researchers to consider more complex resampling techniques that are more appropriate to the type of data and type of analysis that they might run.

References

Bollen, K. A., & Stine, R. (1993). Bootstrapping goodness-of-fit measures in structural equation models. In K. A. Bollen & J. S. Long, (Eds.), *Testing structural equation models* (pp. 111-135). Newbury Park, CA: Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: SAGE publications.

Carver, R.P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Dalgleish, L. I. (1994). Discriminant analysis: Statistical inference using the jackknife and bootstrap procedures. *Psychological Bulletin*, *116*, 498-508.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, May, 116-130.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, *72*, 45-48.

Fan, X., & Jacoby, W. R. (1995). BOOTSREG: A SAS matrix language program for bootstrapping linear regression models. *Educational and Psychological Measurement*, *55*, 764-768.

Fan, X., & Mathews, T. A. (1994, April). *Using bootstrap procedures to assess the issue of predictive bias in college GPA prediction for ethnic groups*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No: ED 372 117)

Fan, X. & Wang, L. (1996). Comparability of jackknife and bootstrap results: An investigation for a case of canonical analysis. *Journal of Experimental Education*, *64*, 173-189.

Harris, D. J., & Kolen, M. J. (1989). Examining the stability of Angoff's delta item bias statistic using bootstrap. *Educational and Psychological Measurement*, *49*, 81-87.

Kotz, S., & Johnson, N. L. (1992). *Breakthroughs in statistics: Volumes 1 and 2*. New York: Springer-Verlag.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Lambert, Z. V., Wildt, A. R., & Durand, R. M. (1991). Approximating confidence interval for factor loadings. *Multivariate Behavioral Research*, *26*, 421-434.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. Pacific Grove, CA: Duxbury.

Mendoza, J. L., Hart, D. E., & Powell, A. (1991). A bootstrap confidence interval based on a correlation corrected for range restriction. *Multivariate Behavioral Research*, *26*, 255-269.

Rasmussen, J. L. (1987). Estimating the correlation coefficients: Bootstrap and parametric approaches. *Psychological Bulletin*, *101*, 136-139.

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities: A Contemporary Journal*, *2*(1), 30-38.

Shaver, J.P. (1993). What significance testing is, and what it isn't. *Journal of Experimental Education*, *61*, 293-316.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *24*, 323-355.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Stine, R. A. (1989). An introduction to bootstrap methods: Examples and ideas. *Sociological Methods and Research*, *8*, 243-291.

Thompson, B. (1988). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. *Educational and Psychological Measurement*, *48*, 681-686.

Thompson, B. (1992). DISCSTRA: A computer program that computes bootstrap resampling estimates of descriptive discriminant analysis function structure coefficients and group centroids. *Educational and Psychological Measurement*, *52*, 905-911.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361-377.

Thompson, B. (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. *Educational and Psychological Measurement*, *55*, 84-94.

Yung, Y., & Bentler, P. M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 195-226). Mahwah, New Jersey: Lawrence Erlbaum.

Send correspondence to: J. Kyle Roberts, Ph.D.
P.O. Box 311335
University of North Texas
Denton, TX 76203-1335
Email: kroberts@unt.edu

Appendix A

SAS Code for the Nested Bootstrap

```
*** SAS PROGRAM FOR BOOTSTRAPPING INDIVIDUALS WITHIN EACH SCHOOL ***;

LIBNAME BTHSB20 'C:\HLM Bootstrap';

DATA HSB20; INFILE 'C:\HLM Bootstrap\HSB20.TXT';
  INPUT SCHID MATH SECTOR CSES;

     *** direct the SAS log to a disk file to avoid SAS LOG Window becoming full;
PROC PRINTTO LOG='C:\HLM Bootstrap\LOGFILE.TMP';
 RUN;

%MACRO BTRAP;           *** start of bootstrap macro 'BTRAP';
%DO BTRAP=1 %TO 2000;   *** 2000 bootstrapped samples, about 4.5 sec. each iteration;
%DO A=1 %TO 20;         *** select each school sequentially;
DATA D1; SET HSB20;     *** (20 schools in the data set, unequal N in each school);
   IF SCHID=&A;

     *** sampling with replacement within each selected school;
     *** bootstrapped sample size equal to the original sample size in each school;
     *** bootstrapped sample within each school is named BTDATA_n;
 DATA BTDATA;
  DROP I;
  DO I=1 TO N;
    IOBS=INT(RANUNI(0)*N) + 1;
    SET D1 POINT=IOBS NOBS=N;
  OUTPUT;
  END;
STOP;
      *** assign a unique random number for later combining data sets;
DATA BTDATA_&A;
  SET BTDATA; UNIQUE=RANNOR(0);

%IF &A=1 %THEN %DO;
      DATA BTDATA_ALL; SET BTDATA_&A;
    %END;
    %IF &A>1 %THEN %DO;
       PROC SORT DATA=BTDATA_ALL;   BY UNIQUE; RUN;
      PROC SORT DATA=BTDATA_&A;    BY UNIQUE; RUN;

          *** combining bootstrapped samples from each school;
      DATA BTDATA_ALL;
      UPDATE BTDATA_ALL BTDATA_&A;  BY UNIQUE; RUN;
    %END;
%END;
```

```
   *** direct PROC MIXED output to a file on disk;
   *** avoids potential problem of SAS Output Window becoming too full;
  FILENAME MIXEDOUT 'C:\HLM Bootstrap\MIXEDFILE';
PROC PRINTTO PRINT=MIXEDOUT NEW;
RUN;

PROC MIXED data=btdata_all NOCLPRINT COVTEST NOITPRINT;
  CLASS SCHID;
  MODEL MATH = SECTOR CSES SECTOR*CSES/SOLUTION DDFM=BW NOTEST;
  RANDOM INTERCEPT CSES/TYPE=UN SUB=SCHID;
     ODS OUTPUT COVPARMS=CP;      *** output random cov. terms to a SAS-DATA-SET;
     ODS OUTPUT SolutionF=FIXED;  *** output fixed effects to a SAS-DATA-SET;
  RUN;

    *** re-direct the output to SAS output window;
PROC PRINTTO PRINT=PRINT; RUN;

DATA COV; SET CP;
  KEEP CovParm Estimate;
PROC TRANSPOSE DATA=COV OUT=COVOUT LET; RUN;

      *** obtain the variances/covariance of random effects;
      *** Use Bryk and Raudenbush notations;
DATA COVOUT; SET COVOUT;
  DROP _NAME_; RENAME COL1=U0 COL2=U01 COL3=U1 COL4=R;

DATA COEFF; SET FIXED;
  KEEP EFFECT ESTIMATE;
PROC TRANSPOSE DATA=COEFF OUT=COEFF LET; RUN;

     *** obtain the model parameter estimates of the fixed effects;
     *** Use Bryk and Raudenbush notations;
DATA COEFF; SET COEFF;
  DROP _NAME_; RENAME COL1=GA00 COL2=GA01 COL3=GA10 COL4=GA11;

     *** combine the two data sets to have one observation for each bootstrapped
sample;
DATA BOTH; MERGE COVOUT COEFF;

      *** append estimates from each bootstrap iteration;
      *** to a permanent SAS dataset on disk: HSB20_L1ONLY;
PROC APPEND BASE=BTHSB20.HSB20_L1ONLY FORCE; RUN;
%END;                       *** end bootstrap iterations;
%MEND BTRAP;                *** end of bootstrap macro;
%BTRAP;                     *** execute the BTRAP macro;


/*
    *** read in the data of bootstrapped results;
    *** (2000 observations from 2000 bootstrap iterations);

DATA TEMP;
  SET BTHSB20.HSB20_L1ONLY;

    *** obtain some basic descriptive statistics for;
    *** the bootstrapped distributions of the estimates;

PROC means n mean std skew kurtosis min max maxdec=3;
  title 'descriptive statistics of HLM model - HSB data';
  title2 'bootstrap individuals within each school';
RUN;
*/
```

Appendix B

S-PLUS Code for the Nested Bootstrap

```
### Nested Bootstrap for HLM

### The following code is for performing a nested bootstrap within
### the HLM framework using lme in S-PLUS.

### Identify the number of schools or groups here
schools<-c(20)

### In the split command, identify the dataset and then the grouping variable
abc<-split(Hsb20, Hsb20$schid)

### Identify the number of bootstrap samples to be drawn
nboot<-2000

### In this matrix, the number of columns must equal number
### of components to be extracted
results.out<-matrix(0, ncol=8, nrow=nboot)
for (j in 1:nboot){

abc.total<-abc[[1]][1,]
for(i in 1:schools){
  data.index <- sample(nrow(abc[[i]]), size =
                 nrow(abc[[i]]), replace = T)
  temp<-abc[[i]][data.index,]
  abc.total<-rbind(abc.total,temp)}

  abc.total<-abc.total[2:nrow(abc.total),]
  final.data<-data.frame(abc.total)

### Define the lme model here but note that the dataset in this
### case is final.data and not your original dataset
  model.out<-menuLme(fixed = math~sector*cses, data = final.data,
        random = ~  cses  |  schid, method = "ML")

### Define which components you want to extract from lme here
results.out[j,]<-c(VarCorr(model.out)[1], VarCorr(model.out)[8],
            VarCorr(model.out)[2], VarCorr(model.out)[3],
            model.out$coefficients$fixed[1],
            model.out$coefficients$fixed[2], model.out$coefficients$fixed[3],
            model.out$coefficients$fixed[4])}
```

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:
Jeffrey B. Hecht, MLR:GLM/SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854.
jbhecht@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*.

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*.

EDITORS

T. Mark Beasley, Editor *MLRV*
Department of Biostatistics
School of Public Health
343C Ryals Public Health Bldg.
University of Alabama at Birmingham
Birmingham, AL 35294
(205) 975-4957 (voice)
(205) 975-2540 (fax)
mbeasley@uab.edu

Robin K. Henson, Associate Editor
Department of Technology and Cognition
College of Education,
P.O. Box 311337
University of North Texas
Denton, Texas 76203-1337
(940) 369-8385 (voice)
(940) 565-2185 (fax)
rhenson@tac.coe.unt.edu

ORDER INFORMATION

Jeffrey B. Hecht, MLR/GLM SIG Executive Secretary
Department of Educational Technology, Research & Assessment
Northern Illinois University
DeKalb, IL 60115-2854
jbhecht@niu.edu

Check out our website at: http://www.coe.unt.edu/mlrv/

POSTMASTER: Send address changes to:
Jeffrey B. Hecht, 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**.