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 (MLR:GLM/SIG)

MLRV

Volume 23 • Number 1 • Fall 1996

Table of Contents

Editorial Board

Multiple Linear Regression Viewpoints (ISSN 0195-7171) is published by the AERA Special Interest Group on Multiple Linear Regression: the General Linear Model (MLR:GLM/SIG) through the University of Missouri at St. Louis. *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* and with the *FAXON* and *READMORE* subscription agencies. MLR:GLM/SIG information and a membership application form can be obtained by writing, FAXing (314-516-5784), Voice Mailing (314-516-5785), or e-MAILing (sspaner@umslvma.umsl.edu) the Executive Secretary. 1996-97 SIG membership and subscription fees are: Individual - $10 for one year, $18 for two years; Library/Agency - $20 per year; and Student - $5 for one year. Fee payment should be made payable to the **Multiple Linear Regression SIG** and sent to Steven Spaner, MLR:GLM/SIG Executive Secretary, Department of Behavioral Studies, University of Missouri - St. Louis, 8001 Natural Bridge Road, St. Louis, MO 63121-4499.

Empirical Characteristics of Centering Methods for Level-1 Predictor Variables in HLM

Randall E. Schumacker, Ph.D.
University of North Texas
Karen Bembry
Dallas Public Schools

Research has suggested that important research questions can be addressed with meaningful interpretations using hierarchical linear modeling. The proper interpretation of results, however, is invariably linked to the choice of centering for the Level-1 predictor variables which produce the outcome measures for the Level-2 regression analysis. In this study, three centering methods (uncentered, group mean, and grand mean) were compared using Read93 and Lunch Status as Level-1 predictor variables of ITBS94 reading test scores. The reliability estimates, or how accurately the sample estimate represents the population value, differed among the three centering methods. It was found that the group mean centering method provided the better reliability estimate. When using outcome measures based upon these three centering methods in a Level-2 analysis using two predictors, Gradrate and Percent Advdip, the group mean centering method indicated a more reliable estimate, but the grand mean centering method explained more between school variance. In fact, the gamma regression coefficients were markedly different, and the amount of variance explained was no longer consistent across the centering methods. These findings indicate that the choice of centering method for Level-1 predictor variables can affect empirical findings in HLM.

In quantitative research, it is essential that the variables under study are meaningful and interpretable so that statistical results can be related to theoretical concerns (Arnold, 1992). This principle is especially meaningful in multi-level analyses of variables such as in hierarchical linear modeling (HLM). In hierarchical linear modeling, the Level-1 variable s intercepts and slopes become outcome variables for Level-2 analyses. Because of potentially complex "nested" designs, it is important that each variables' value be clearly understood and specifically articulated (Bryk & Raudenbush, 1992).

Hierarchical linear modeling can be used to investigate many of the research questions in education that involve at least two levels of variables. Samples of such questions include: Do schools with a high percentage of students with limited English proficiency also have high achievement scores? Is the relationship between student SES and achievement invariant across schools? In fact, several studies investigating teacher effectiveness, school effectiveness, and student change and growth have been conducted using HLM (Bryk & Raudenbush, 1987 & 1988, Raudenbush, 1988, Lee & Bryk, 1989, Mendro et al. 1994, Webster et al, 1994). These studies recognize the nested design structure of students within classrooms, classrooms within schools, and schools within districts which produce different variance components for variables at each level.

In multi-level analyses, variables measured at the different levels provide different variance estimates (Bock, 1989), and depending on how the data are treated, opposing conclusions can be reached (Kreft, 1995; Kreft, de Leeuw, Aiken, 1995). For example, school level variables do not vary for students in a particular school. These school-level variables instead help to explain between-school variance rather than within-school variance. Likewise, students in the same classroom or school tend to be more alike than in other classrooms or schools; hence, the variance between students is not constant. Similarly, interpretations of outcomes can vary at the school-level, often leading to conflicting results. Student level data, however, measures the within-school variance, conditioned by school-level effects. In other words, the scores of students in each school building are adjusted using school-level variables, such as the crowded condition of that campus, to better reflect the nature and interpretation of the scores. A typical research question for an HLM analysis would be the investigation of the effect of a school s graduation rate and percent of students in advanced diploma plans on the mean reading test scores of 9th graders. In HLM terminology, this is a "means as outcomes" approach which involves an examination and use of the intercept values as outcomes (dependent variable) for Level-2 variable analysis. The ability to statistically analyze these characteristics within each school, until recently, has been overlooked. Most data analyses have been done using multiple regression single-level variable models.

One critical aspect to conducting HLM

analyses is centering Level-1 predictor variables that produce the outcomes that are used as dependent variables in Level-2 analyses. The interpretation of these outcomes is critical to the meaningfulness of results since centering changes, not only the coefficient s value, but also the research questions being answered by the statistical analysis (Burton, 1993). Theory should drive the decision to center any Level-1 variable as indicated by the research questions included in the investigation. This policy is in keeping with appropriate multiple linear regression procedures. With the introduction of HLM, however, the effect of one level of variables on another introduces several areas for further investigation (see conclusion section). The focus of this paper is on one such area, namely, centering effects of Level-1 variables.

Four possibilities exist for centering Level-1 predictor variables in HLM: X metric, grand mean, group mean, and user defined location, such as a cut-off score (Bryk & Raudenbush 1992). This study included the first three centering methods to determine whether the Level-1 centering decision affects the reliability estimates in the HLM analysis. This investigation further examined how centering decisions made for Level 1 variables affect the amount of between-school variance explained by Level-2 variables.

METHOD

Data Set

Research questions posed for this study were investigated using data from ninth grade students (n = 5638) continuously enrolled in 26 high schools within a large urban school district. The Level-1 variables in this study were defined as student level variables. The student level variables selected for this study included individual reading test scores from the Iowa Test of Basic Skills (ITBS94) for 1994 as the dependent variable. The 1993 individual reading scores (Read93) and an individual student socio-economic indicator identifying free-lunch status (Lunch Status) were the two independent predictor variables. The reading test scores were interval level data with a potential range from 0 to 26. Lunch Status was a dichotomous variable indicating whether or not a student was in the free lunch program.

Level-2 variables were defined as school-level variables. School level variables from the twenty-six high schools selected were the graduation rate for each high school (Gradrate) and the percent of the students in advanced diploma plans within each school (%Advdip). No Level-2 variables used in the study were aggregates of any individual Level-1 variables. Only the effects of the centering options on the "means as outcomes" or the intercept was investigated in this study.

Research Questions

Prior research has indicated that both an interpretation of intercept outcome values and a change in the research question occurs based upon a choice of centering method. Our concern, therefore, was not with theoretical issues which should be answered as an aspect of the research design, but with the empirical issues surrounding the reliability estimates. Thee reliability estimates represent how well the sample mean reflects the population mean and whether the amount of between-school variance predicted at Level-2 would be the same.

Analyses

Several analyses specifying different models were undertaken to answer the research questions. An initial analysis established a "fully unconditional" model, or a model without any Level 1 or Level 2 predictors (Bryk & Raudenbush, 1992). Two separate models with only a single Level 1 predictor variable were then specified. This was followed by a two predictor model with both variables included. A final analysis included a model with both Level-1 predictors (READ93, LUNCH) and two Level-2 predictors (AdvDip, Gradrate). Three analyses were run on each of these models. The analyses involved either an uncentered predictor, a predictor centered on the grand mean, or a predictor centered on the group mean. The Level-two predictors were not centered. The models are specified next.

Fully Unconditional Model

Student level (Level 1) $Y_{ij} = \beta_{oj} + r_{ij}$

where

Y_{ij} = ITBS 94 reading score for student I in school j

β_{oj} = mean reading score in school j

r_{ij} = Level-1 error, $N(0, \sigma^2)$; σ^2 = student level variance

School level (Level 2) $\beta_{oj} = \gamma_{oo} + u_{oj}$

where

β_{oj} = mean reading score in school j

γ_{oo} = grand mean of the district (N=26 schools)

u_{oj} = random effect school j, $N(0, \tau_{oo})$; τ_{oo} = school level variance

Level 1 Predictor Models

READ93 model

$Y_{ij} = \beta_{0j} + \beta_{1j}(READ93) + r_{ij}$

where
Y_{ij} = ITBS 94 reading score for student I in school j
β_{0j} = mean for school j
β_{1j} = slope for school j
r_{ij} = Level-1 error

$\beta_{0j} = \gamma_{00} + u_{0j}$

where
γ_{00} = intercept mean of the district (n=26 schools)
u_{0j} = random effect for school j

$\beta_{1j} = \gamma_{10} + u_{1j}$

where
γ_{10} = slope mean of the district (n=26 schools)
u_{1j} = random effect for school j

Lunch model

$Y_{ij} = \beta_{0j} + \beta_{1j}(LUNCH) + r_{ij}$

where
Y_{ij} = ITBS 94 reading score for student I in school j
β_{0j} = mean for school j
β_{1j} = slope for school j
r_{ij} = Level-1 error

$\beta_{0j} = \gamma_{00} + u_{0j}$

where
γ_{00} = intercept mean of the district (n=26 schools)
u_{0j} = random effect for school j

$\beta_{1j} = \gamma_{10} + u_{1j}$

where
γ_{10} = slope mean of the district (n=26 schools)
u_{1j} = random effect for school j

Read93 and Lunch model

$Y_{ij} = \beta_{0j} + \beta_{1j}(READ93) + \beta_{2j}(LUNCH) + r_{ij}$

where
Y_{ij} = ITBS 94 reading score for student I in school j
β_{0j} = intercept for school j
β_{1j} = slope of READ93 for school j
β_{2j} = slope of LUNCH for school j
r_{ij} = Level-1 error

$\beta_{0j} = \gamma_{00} + u_{0j}$

where
γ_{00} = slope mean of the district (n=26 schools)
u_{0j} = random effect for school j

$\beta_{1j} = \gamma_{10} + u_{1j}$

where
γ_{10} = READ93 mean slope in the district
u_{1j} = random effect for school j

$\beta_{2j} = \gamma_{20} + u_{2j}$

where
γ_{20} = Lunch mean slope in the district
u_{2j} = random effect for school j

Level 1 and Level 2 Predictor Models

$Y_{ij} = \beta_{0j} + \beta_{1j}(READ93) + \beta_{2j}(LUNCH) + r_{ij}$

where
Y_{ij} = ITBS 94 reading score for student I in school j
β_{0j} = mean for school j
r_{ij} = Level-1 error

$\beta_{0j} = \gamma_{00} + \gamma_{01}(AdvDip) + \gamma_{02}(Gradrate) + u_{0j}$

where
γ_{00} = intercept mean of the district (n=26 schools)
u_{0j} = random effect for school j

$\beta_{1j} = \gamma_{10} + \gamma_{11}(AdvDip) + \gamma_{12}(Gradrate) + u_{1j}$

where
γ_{10} = Read93 slope mean of the district
u_{1j} = random effect for school j

$\beta_{2j} = \gamma_{20} + \gamma_{21}(AdvDip) + \gamma_{22}(Gradrate) + u_{2j}$

where
γ_{20} = Lunch slope mean of the district
u_{2j} = random effect for school j

The combined equation for the full model with both Level 1 and Level 2 predictor variables is then specified as:

$$Y_{ij} = (\gamma_{00}+\gamma_{01}(\text{AdvDip})+\gamma_{02}(\text{Gradrate})+u_{0j}) + (\gamma_{10}+\gamma_{11}(\text{AdvDip}) + \gamma_{12}(\text{Gradrate})+u_{1j})(\text{READ93}) + (\gamma_{20}+\gamma_{21}(\text{AdvDip}) + \gamma_{22}(\text{Gradrate})+u_{2j})(\text{LUNCH}) + r_{ij}$$

RESULTS

Level 1 Variable Analyses

The "fully unconditional" model, which only specified an intercept, resulted in a reliability estimate of .98 (Table 1). This initial "fully unconditional" null model allows us to partition the total variance in reading scores into a between school variance component (24%). It also establishes an estimate for the grand mean (β_0) , a confidence interval (β_0 +/- $SE\beta_0$), and establishes the parameters for within-school variability (^{2}) and between school variability ($_{00}$). The reliability estimate indicates how well each school s sample average in reading achievement estimates their true mean (Bryk & Raudenbush, 1992). In this case, the reliability estimate was .98, indicating that the school s sample means are quite reliable as indicators of their true school means. The significant t-value indicates that the schools do not have the same mean ITBS 1994 reading average.

In the single Level 1 predictor model for READ93, results indicated that the reliability estimates differed between the three centering methods. The slope and reliability estimate, however, were the same as in the "fully unconditional" model. As expected, the amount of within school variance remained the same regardless of which centering method was used (45%).

Table 3 indicates results from the three centering methods when using Lunch as a single Level-1 predictor variable. The group mean centering method yielded results identical to the "fully unconditional" model, and the grand mean centering method more closely approximated this initial model than the uncentered approach. The amount of within school variance explained was small (3%), and as expected, the same regardless of choice of centering method.

Table 4 lists the results of the "fully unconditional" analysis and further indicates the effect of each centering method w n both Read93 and Lunch Status were used in a Level-1 prediction equation for 1994 ITBS reading outcomes. The results indicated that 45% of the between school variance was explained when using both predictors, which was the same amount indicated when using Re 93 alone, suggesting that the Lunch variable doesn't contribute any additional explained variance in the model. Moreover, the sample mean intercept value using the group mean centering method was the same as in the initial "fully unconditional" model, with only a slight improvement in the reliability estimate (.98 to .99). The reliability estimate for the grand mean centering method was more approximate to these values than the uncentered method, especially when using Read93 as the only predictor. The group mean centering method was therefore the most stable of the three centering methods.

From a practical research point of view, the choice of Level-1 predictors will impact the amount of within school variance explained. In our approach, preference would be given to using only Read93 as a Level-1 predictor since Lunch did not add any additional significant variance explained. However, for our purposes, we continued to use both Level-1 predictor variables in the Level-2 equation.

Level 1 and Level 2 variable analysis

Table 5 indicates each type of centering method and the associated summary statistics from the Level 2 complete model prediction equation. The amount of between-school variance explained is no longer consistent across the centering methods. The amount of variance explained using uncentered Level-1 variables was 89%; with group mean centering it was 70 %; and with grand mean centering it was 92%. The reliability estimates, or how well the sample estimates indicate the true population values, also differed. The group mean centering method yielded the highest reliability estimate (.96), but indicated very different coefficients for the variables than the other two centering methods, and had the lowest percent variance explained (70%). This leads to conflicting results since the group mean centering method was preferred in the Level 1 analyses, but the grand mean centering method explained more between-school variance in the Level 2 analysis.

Table 1. Full Unconditional Model on 1994 ITBS reading scores (n=26 schools).

Centering Method	β_0	SE β_0	r_{xx}
Null model	16.85	.67	.98

Note: No predictors were specified in Level 1 analysis; Intraclass correlation coefficient = $\tau_{00}/(\tau_{00} + {}^{2}) = 11.60/(11.60 + 36.08) = 24\%$ of variance in 1994 ITBS reading scores explained between schools; and regression coefficient is significant (critical t =25.15, p=.0001).

Table 2. Level 1 predictor READ93 on 1994 ITBS reading scores (n=26 schools).

Centering Method[a]	β_0	SE β_0	r_{xx}	β_1	SE β_1	r_{xx}
Uncentered	13.85	.55	.69	1.87	.33	.67
Centered: Group Mean	16.85	.67	.98	1.82	.33	.67
Centered: Grand Mean	16.74	.62	.97	1.87	.33	.67

a Intraclass correlation the same for each centering method [σ^2 (ANOVA) - σ^2 (READ93)/ σ^2 (ANOVA) = 36.08-20.00/36.08 = 45%

Table 3. Level 1 predictor Lunch on 1994 ITBS reading scores (n=26 schools).

Centering Method[a]	β_0	SEβ_0	r_{xx}	β_1	SE β_1	r_{xx}
Uncentered	13.85	.55	.69	1.87	.33	.67
Centered: Group Mean	16.85	.67	.98	1.82	.33	.67
Centered: Grand Mean	16.74	.62	.97	1.87	.33	.67

a Intraclass correlation the same for each centering method [σ^2 (ANOVA) - σ^2 (Lunch)/ σ^2 (ANOVA) = 36.08-35.05/36.08 = 3%

Table 4. Both Level 1 predictor variables on 1994 ITBS reading scores (n=26 schools).

	Unconditional			Read93			Lunch		
Centering Method[a]	β_0	SE β_0	r_{xx}	β_1	SE β_1	r_{xx}	β_2	βSE $_2$	r_{xx}
Uncentered	6.46	.34	.47	.22	.004	.30	.69	.15	.25
Centered: Group Mean	16.85	.68	.99	.22	.004	.37	.69	.16	.28
Centered: Grand Mean	16.77	.26	.89	.22	.004	.29	.69	.15	.26

a Intraclass correlation the same for each centering method [σ^2 (ANOVA) - σ^2(READ93 & Lunch)/ σ^2(ANOVA) = 36.08-19.88/36.08 = 45%

Table 5. Complete model: Graduation Rate and Percent Advanced Diploma using 1994 ITBS READ93 and Lunch reading score intercepts (n = 26 schools).

Centering Method	γ_{00}	SE γ_{00}	γ_{01}	SE γ_{01}	γ_{02}	SE γ_{02}	r_{xx}
Uncentered[a]	4.17	1.11	.003	.02	.04	.03	.43 3.75
Centered: Group Mean[b]	8.76	1.19	.030	.02	.13	.03	.96
Centered: Grand Mean[c]	14.15	.68	.004	.01	.05	.02	.84

[a] Intraclass correlation coefficient = τ_{00} (ANOVA) - τ_{00} (Gradrate & %Advdip)/ τ_{00}(ANOVA) = 11.60 - 1.22/ 11.60 = 89%; t = 3.75, p > .002.

[b] Intraclass correlation coefficient = τ_{00} (ANOVA) - τ_{00} (Gradrate & %Advdip)/ τ_{00}(ANOVA) = 11.60 - 3.46/ 11.60 = 70%; t = 7.36, p > .0001.

[c] Intraclass correlation coefficient = $_{00}$ (ANOVA) - $_{00}$ (Gradrate & %Advdip)/ τ_{00}(ANOVA) = 11.60 - .97/ 11.60 = 92%; t = 20.87, p > .00001.

CONCLUSIONS AND DISCUSSION

In practical applications, Level 1 predictor variables appear to become more stable when they are centered on either the group mean or grand mean. In our study, the initial sample estimate (intercept, β_0) was close to the population value in the "fully unconditional" model, as indicated by the reliability estimate of .98. This finding is expected in any initial null model. The reliability estimates, however, differed between the three centering methods when centering the Level 1 predictors Read93 and Lunch. For Read93, the reliability estimates were .76 uncentered), .98 (group-mean centered), and .90 (grand-mean centered). For Lunch, the reliability estimates were .69 (uncentered), .98 (group-mean centered), and .97 (grand-mean centered). The group-mean centering method for both Level 1 predictor variables yielded the same intercept and reliability estimate as in the "fully unconditional model". The intercept and slope values differed in the grand mean centering and uncentered methods, although they were more approximate when using grand mean centering. As expected, the amount of within school variance explained remained the same regardless of which centering method was used (45%). When using outcome measures based upon these three centering methods in a Level 2 full model analysis with two predictors, Gradrate and Percent Advdip, the group mean centering method also indicated a more reliable estimate, but the grand mean centering method explained more between school variance. The gamma regression coefficients were markedly different and the amount of variance explained was no longer consistent across the centering methods. These findings indicate that the centering of Level 1 variables empirically effects the variance estimation in Level 2 model analyses.

We found that the meaningfulness of the intercept and slope values in a Level 1 (student level) model depends upon the centering of the Level 1 predictor variables. In raw metric form, the equation $Y_{ij} = \beta_{oj} + \beta_1 X_{ij} + r_{ij}$, yields intercept values, β_{0j}, which are interpreted as an outcome measure for a student attending school j who has a 0 (zero) on X_{ij}. Obviously this causes a problem in the interpretation of student achievement using these raw metric intercept values because the lowest score on the test will not be zero. When centering Level 1 predictor variables around the grand mean, they are determined by: $(X_{ij} - X..)$. The intercept, β_{0j}, can now be interpreted as an outcome measure for a student in school j whose value on X_{ij} is referenced to the grand mean. This permits a useful interpretation of the intercept as an adjusted mean for school j: in this case, $\beta_{0j} = \mu_{Yj} + \beta_{1j}(X_{.j} - X..)$. This is similar to the adjusted means in an ANCOVA analysis. These intercept values can now represent a specific interpretation of the outcome measures for each school in the Level 2 model analysis. The intercept variance term reflects the variation in the adjusted means for the set of schools. If the Level 1 predictor variables are centered around the group mean, they are determined by $(X_{ij} - X_{.j})$. Now the intercept, β_{0j}, represents the unadjusted outcome measure for a student in school j. In this instance,

$\beta_{0j} = \mu_{Yj}$. The intercept variance, Var (β_{0j}), is now the variance around the Level 2 variable unit means, μ_{Yj}. This permits an examination of the sampling distribution of the school means or slopes around a population mean value, i.e. district mean value.

A researcher will typically center some or all Level 1 (student-level) predictors at either the grand mean or group mean to add stability to the estimation process and provide for intercepts that can be meaningfully interpreted. Centering, however, also has the effect of changing the coefficients that are estimated and altering the research question(s) being asked. Burton (1993), using a NELS88 data set (outcome =mathematics achievement test; student-level variables =minority status, socio-economic status, and absenteeism; school-level variables =percent minority; location of school, and percent low SES), indicated that uncentered and grand mean centering indicated only significant Level 1 coefficients while group mean centering indicated significant Level 2 coefficients (school level). This implied two different interpretations of results: one at the student level with individual status affecting achievement, and one at the school level with average school status affecting achievement. It is troublesome that a choice between these two centering methods could result in two different interpretations. Which is the correct interpretation of the results?

Research has suggested that important research questions can be addressed with meaningful interpretations using hierarchical linear modeling (Raudenbush, Rowan, & Cheong, 1993). For practical applications, the unconditional model allows partitioning of variance into within-school and between-school components for the outcome measure. The choice of variables at Level-1 impacts the amount of within-school variance (student-level) that can be explained, and the choice of variables at Level-2 impacts the amount of between-school variance (school-level) that can be explained given the outcome measures provided from the Level-1 equation. The proper interpretation of results, however, is invariably linked to the choice of centering for the Level-1 predictor variables which produces the dependent measures for Level-2 regression analyses. Studies which examined organizational level, school effectiveness, and teacher effectiveness variables using hierarchical linear models have provided more appropriate variance estimates and means as outcomes than previous single level data analyses. The proper interpretation and accuracy of estimation, however, requires that a researcher pay special attention to the centering effects in Level-1 student-level variables upon Level-2 analyses when conducting hierarchical linear models.

For many researchers, multiple regression has become a valuable data analytic tool because many of the issues related to using multiple regression have been investigated. For example, sample size and power, non-normality, heterogeneity, number of predictors, ratio of sample size to predictors, multi-collinearity, use of composite variables, outliers, and interaction effects. We believe that many of these concerns need to be restated in the context of hierarchical linear modeling. Once case in point is the effect of centering when including an interaction term. Aiken & West (1993) have indicated that centering variables in the presence of an interaction term in multiple regression changes the value of the regression coefficients. In HLM, this would follow as a dictum, especially in light of the findings by Burton (1993). Additional examination of other factors will determine what effect, if any, they have upon hierarchical linear analyses.

REFERENCES

Aiken, L. S. & West, S. G. (1991). Multiple *Regression: Testing and Interpreting Interactions*. Newbury Park, CA: Sage Publications.

Arnold, C. L. (1992). An introduction to hierarchical linear models. *Measurement and Evaluation in Counseling and Development, 25,* 58-90.

Bock, R. Darrell. (1989). *Multilevel Analysis of Educational Data*. San Diego, CA: Academic Press.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications, and Data Analysis Methods*. Newbury Park, CA: Sage Publications.

Bryk, A. S. & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, November, 65-108.

Bryk, A. S. & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147-158.

Burton, Bob. (1993, April). Some observations on the effect of centering on the results obtained from hierarchical linear modeling. (Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, GA).

Kreft, I.G.G. (1995, April). The effects of centering in multilevel analysis: Is the Public school the loser or the winner? A new analysis of an old question. Paper presented at the American Educational Research Association annual meeting. San Francisco, California.

Kreft, I.G.G., de Leeuw, J., Aiken, L.S. (1995). The effect of different forms of centering in hierarchical linear modeling. Multivariate Behavioral Research, in press.

ee V. E. & Bryk, A. S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education*, 62, 172-192.

Mendro, R. L., Webster, W. J., Bembry, K. L. & Orsak, T. H. (1994, October). Applications of Hierarchical Linear Models in Identifying Effective Schools. (Paper presented at the Annual Meeting of the Rocky Mountain Educational Research Association, Tempe, AZ).

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, 13, 85-116.

Webster, W. J., Mendro, R. L., Bembry, K. L. & Orsak, T. H. (1994, October). Alternative methodologies for identifying effective schools (Paper presented at the Annual Meeting of the Rocky Mountain Educational Research Association, October, Tempe, AZ).

Comments On Validation Methods For Two Group Classification Models Widely Accepted In Credit Scoring Or Response Analysis

Timothy H. Lee,
Equifax Decision Systems,
Equifax Credit Information Services, Inc.,
Internal Mail Code 42S,
P.O. Box 740006, Atlanta, Ga 30374-0006

The two group classification methods are popular approaches for the separation of one group from the other. For these purposes either parametric or non-parametric classification approaches are used. In many cases a scoring algorithm is derived and the score distribution serves as a basis of the decision making. Generally, validation of a model is to assure the model has reasonable separation power when it is applied to a different data set not used for the development of the model, i.e., holdout data set. In the credit scoring case, Regulation B of Equal Credit Opportunity Act requires the scoring algorithm be revalidated frequently enough to ensure that it continues to meet statistical standards. In addition, in case of comparison of more than one model, it is necessary to quantify model performance in some way. Two sample Kolmogorov-Smirnov test statistic, Kullback-Leibler Number, and Mahalanobis Distance, etc. are popular ways of quantifying model performance. In this study, such popular methods are discussed along with the advantages and disadvantages of each method using a simulated data set and a suggestion of an improved, intuitive, and simple quantifying method for model performance is made

KEY WORDS: Kullback-Leibler number, Two sample Kolmogorov-Smirnov Test, Logistic Regression, Discriminant Function

1. INTRODUCTION.

Two group classification analysis is a very popular approach in industries such as credit granting or target marketing. For instance, in the credit industry, credit grantors want to predict the creditworthiness of applicants. By the two group classification approach, more creditworthy applicants are separated from less creditworthy applicants. In the process of discrimination, a scoring algorithm is derived based on the known data and the algorithm is applied to applicants to score them. Without a doubt, a good scoring algorithm has better separation power than others. Of course, the ultimate performance of the model should be measured by the profitability. The profitability, however, is hard to be measured objectively. Besides, there are various uncontrollable econo-socio, consumer behavior related, or business related factors that affect profitability. In this paper, we would like to focus our attention on the separation power and separation pattern of a classification of models - especially on the quantification of model performance.

The measurement of model performance is important to see if the algorithms are discriminating adequately or to determine if other models do a better job of rank ordering. If a model is to screen a potentially better creditworthy applicant for credit extension, Regulation B of the Equal Credit Opportunity Act (ECOA) requires frequent validation of the model performance. It, however, does not point to any specific statistical method. The regulation simply states, 'The scoring system must be periodically revalidated by the use of appropriate statistical principles and methodology ...' Besides the regulatory reasons, there are many other reasons to quantify the performance of models.

In most applications, the models are for two group classification. The model provides a basis to assign an object to either of the two populations, p_1 or p_2. In the process of classification, multivariate observations $\mathbf{x}$ for each object were transformed to univariate observation y such that the y's derived from populations p_1 and p_2 were separated as much as possible. In the industry, each element in $\mathbf{x}$ is

demographic, socio-econo, or credit bureau factor pertaining to each individual and computed y is called score. The score per each individual is a base for the classification.

Next, we would like to review the statistical validation approaches widely used in the industry and discuss related issues. Finally an alternative measure will be proposed.

2. MEASUREMENT OF SEPARATION.

It has been an issue among analysts using two group classification methods including logistic regression, discriminant function or regression with a binary dependent variable what statistical method will be used to measure the performance of the model. Since most applications are two group classification, the model performance is measured by the accuracy of separation of a group from the other. If a non-parametric classification method is used, the classification error rate would be considered as a measure. In many cases, a parametric or semi-parametric approach is used for classification and score for each individual is computed as a basis for class assignment. In such a case, model performance should not be measured simply by the classification error. The separation pattern of a model should be taken into consideration because it may affect stability of decision.

The score distribution generated for each group differs from each other if the scoring algorithm separates. The degree of difference between the score distributions does not necessarily measure the performance of a model. We will visit this issue in the discussion session again.

Two most commonly used statistical methods for the validation of a model or for the comparison of model performance are i) Two Sample Kolmogorov Smirnov (K-S hereafter) test or ii) computation of the Kullback Leibler Entropy (Divergence) on the score distributions generated by the scoring algorithm. The scoring algorithm, sometimes called a scoring model, is an equation or a rule for assignment derived using any two group classification method such as Discriminant Function, Logistic regression, or other parametric or non-parametric classification technique, etc.

2.1 Two Sample K-S Test

The test was proposed by Smirnov, N. V. (1939) for the test of the hypothesis that any two samples are from the same population. It is to test H_0: $F(x) = G(x)$ for all x against the general alternative H_1 when the two samples, $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ are independent random samples from continuous distributions with c.d.f.'s $F(x)$ and $G(y)$. The test rejects H_0 if and only if the observed value of

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)| \text{ for all } x,$$

where $F_m(x)$ and $G_n(x)$ denote the empirical distributions corresponding to $F(x)$ and $G(x)$, is greater than any threshold value determined at a proper significance level.

The threshold value is to be determined from the table or, when m and n are greater than 80, approximated by

$$z_a ¥ [(m+n)/(m ¥ n)]^{0.5},$$

where z_a is determined by proper significance level.

It is conventional, somehow, in the industry, that the $D_{m,n}$ is used as a measure of model performance. In other words, the test statistic value for the testing of equality of the two distributions is used for the measure of separation power of a model. We will revisit this issue in the following sections.

2.2 Divergence

As Soofi (1994) pointed out in his recent paper about capturing the intangible concept of information, many statisticians are familiar with the theory of discrimination information. Moreover, quantifying information in some statistical problems has been the highlight among statisticians in the industry. Since most often the purpose of the model is to separate one group from the other, the interest of the analysts is in the entropy of discrimination information. Shannon (1948) developed information entropy for quantifying the expected uncertainty associated with an outcome from a sample from a population that has distribution f. His formula for the entropy was

$$H(x) = -E[f(x) \log f(x)].$$

Kullback and Leibler (1951) generalized above entropy into relative entropy,

$$H(f,g) = \int \log [f(x)/g(x)]\, dF(x), \quad \ldots\ldots (1)$$

where f and g are probability distributions for p_1 and p_2, respectively.

Expression (1) is known as Kullback Leibler entropy, directed divergence, or the relative information of class 1 with respect to class 2. The entropy is not a symmetric function. Jeffrey (1946) considered a symmetric version of this function as a measure of divergence between two distributions with densities f and g,

$$D = H(f,g) + H(g,f) \quad \ldots\ldots (2)$$

The quantity is called Kullback Leibler Cross Entropy, Information Number, or Divergence. Right hand side of (2) can be rewritten as

$$\int \log [f(x)/g(x)]\, dF(x) - \int \log [f(x)/g(x)]\, dG(x)$$

$$= E\,[L(x)|p_1] - E\,[L(x)|p_2], \quad \ldots\ldots (3)$$

where $L(x) = \log [f(x)/g(x)]$.

The divergence is expressed as the difference in means of the two L(x)'s on p_1 and p_2, respectively.

Therrien (1989) showed that the divergence is equal to Mahalanobis distance between the two means when the data has Gaussian distribution and the two covariance matrices are equal. If the measure of divergence is applied to score distribution to see how well the two score distributions differentiate each other, the divergence can be written, under the normality and equal variance and covariance assumption,

$$D = (m_1 - m_2)^2/[(s_1^2 + s_2^2)/2], \quad \ldots\ldots (4)$$

where m_1, s_1^2, m_2, and s_2^2 are means and variances of the score distributions for p_1 and p_2, respectively.

In the above we reviewed two commonly used approaches for model validation in the industry. In the following sections we will discuss advantages and disadvantages of the approaches taking examples and a potentially superior alternative approach will be proposed.

3. EXAMPLE.

In the past major model developers in the credit industry have debated regarding selection of the validation methods and they tried to show that their approach was superior to their competitors'. Strange enough, each of major developers employed one approach.

In this section we will assume several cases of separation pattern and compare the changes of the two approaches, K-S test statistic vs. divergence. We will assume some different patterns of separation depending on the skewness conditions of the two scoring distributions for each group as in the following:

1) Two score distribution curves are normally distributed. (See Figure -1.)

2) Two score distribution curves are inwardly skewed. (See Figure - 2.)

3) Two score distribution curves are outwardly skewed. (See Figure - 3.)

4) One score distribution is nested by the other. (See Figure - 4.)

Figure -1. Two Normal Distributions

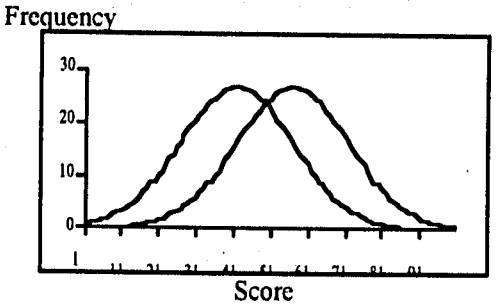


Figure - 2. Two Inwardly Skewed Distributions

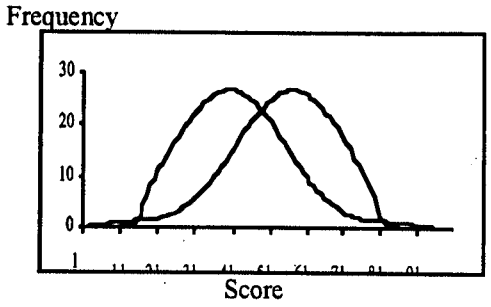


Figure - 3. Two Outwardly Skewed Distributions

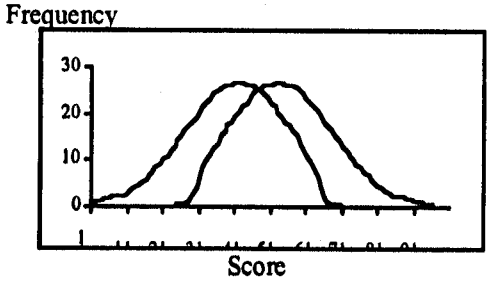
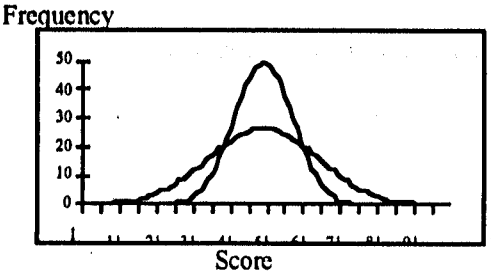


Figure - 4. One Distribution is nested by the other



In case 1) both divergence and K-S test statistic value are used correctly. In case 2) divergence will be measured too conservatively, while in case 3), divergence will give too optimistic measure. In case 4) K-S test statistic value will be little too optimistic but divergence will measure more accurately. Even though most cases are close to the case1), it is a natural desire for the analysts to use a method that measures the separation power of a model properly taking into consideration the separation pattern. In the following section a different idea from the previously mentioned methods will be presented.

4. COEFFICIENT OF SEPARATION.

As mentioned in the previous section, divergence seems to be affected by the skewness of the score distribution, while the Two sample K-S test is not proper as a measure of separation in the case when one distribution is included in the other, even though the test statistic can be a good measure for differentiation of one distribution from the other. To alleviate such problems of the two common measures the following approach is proposed:

1) Create a cumulative empirical score distribution for each group (e.g., creditworthy versus non-creditworthy).

2) Per each observed score point (or interval) read the two cumulative empirical probability as x and y coordinate.

3) Plot the coordinates on the unit square. Then, the trace of the points form a curve reflecting the pattern of separation as in the Figure - 5.

4) Find the area between the curve created in 3) and the 45 degree line (no separation line). If the curve is partially above and below the 45 degree line, find absolute difference of the area below the 45 degree line and that above the line. Such a case is observed when one distribution curve is included inside of the other. (See Figure - 5.)

5) The absolute difference of the areas computed in 4) is divided by the area of the triangle under the 45 degree line. This value will be used as a measure of separation.

Above Procedure is similar to plotting Lorentz curve or ROC (Receiver Operating

Characteristic) curve except finding the difference of the two areas. Simple calculus method such as Trapezoidal approximation of curve would be good enough to estimate the areas. This method (call it Coefficient of separation or C-S) is compared with the two commonly used methods, Two sample K-S test and divergence in Table - 1.

Figure -5. Separation Curve on Unit Square

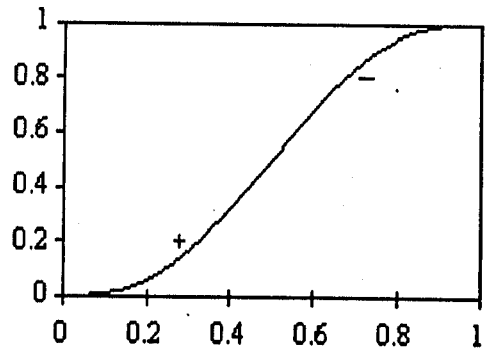


Table - 1. Comparison of model performance measures

Cases	Measures		
	K-S	Divergence	C-S
1. No skewness	38.16	0.99	51.84
2. Skewed inwardly	37.36	0.75	48.32
3. Skewed outwardly	37.65	1.30	55.83
4. One includes the other	14.72	0.00	0.00

5. REMARKS.

In this article we considered three different methods for model validation. It is often observed in the credit industry that selection of a validation method depends on the modeling method. For example, if the modeling approach is parametric or semi-parametric, Two sample K-S test is very often used. If the model is derived by iterative search method maximizing Information number, the measure for model performance is usually divergence. In most cases each of the three method works properly. Extreme cases such as mentioned ahead are very rare, even unrealistic. Such cases, however, can be artificially created by some transformation such as Logistic function. Two sample K-S test statistic value is not affected by any one to one transformation. The divergence, however, is affected when the skewness is changed. The coefficient of separation, compared to the other two methods, seems to be reasonable in most cases as a measure for model performance because it reflects separation pattern of a model.

ACKNOWLEDGMENT

The author thanks Dr. Donald Searls of University of Northern Colorado for his helpful comments and Dr. Ming Zhang of Equifax decision Systems for his careful review of the paper. The author also thanks to Dr. John Pohlmann of Southern Illinois University for helpful suggestions.

REFERENCES

Kullback, S., and Leibler R. A. (1951), "On Information and Sufficiency", *The Annals of Mathematical Statistics*, 22, 986-1005

Smirnov, N. V. (1939), "On the estimation of the discrepancy between empirical curves of distribution for two independent samples.". (Russian) Bull. Moscow Univ. 2, 3-16

Soofi, Ehsan S. (1994), "Capturing the Intangible Concept of Information", *Journal of the American Statistical Association*, 89, 1243-1254

Therrien, Charles W. (1989), *Decision, Estimation, and Classification*, New York: John Wiley

Gender Discrimination Determination in Faculty Salary Patterns from Small-Population Colleges

Sandra Lebsack
University of Nebraska at Kearney
Robert Heiny, Don Searls, Ann Thomas, and John Cooney
University of Northern Colorado

Attempts were made to develop multiple linear regression models to represent salary patterns from two small-population (N=91, N=44) colleges. Multiple discriminant, canonical, and set correlation analyses were used to confirm the presence or absence of "tainted" variables. Problems with multicollinearity were solved by removing variables. "Fixed" models were formulated after using variable selection techniques to determine statistical significance. Entry salary (which acted as a suppressor variable) did not have a linear relationship to salary and the models involving it violated the normality of error terms assumption. Average percent increase in salary was used instead. However, the presence of heteroscedasticity in models for both colleges could not be eliminated. For these colleges, using multiple linear regression to determine, statistically, the presence or absence of gender discrimination in salary patterns was not possible.

In order to detect discrimination in salary based on factors such as sex, race, or ethnic group, comparisons have been made between the discriminatory groups and white males. Mean and median salaries have been used to show overall inconsistencies in salary allocation (Boyd, 1979). Some nondiscriminatory factors have been accepted as reasons for differential salaries. With regard to discrimination due to sex, Greenfield (1977) states that these factors are merit, quality or quantity of production, seniority, or "any reasonable factor other than sex" (p. 43).

Multiple linear regression has been accepted by the legal system for displaying or refuting discrimination (Finkelstein & Levin, 1990; Baldus & Cole, 1980). Some researchers (Hengstler, Muffo & Hengstler, 1982) think it is possibly "the most effective method for analyzing sex discrimination in faculty salaries" (p. 16). Others have also used canonical analysis and multiple discriminant analysis (Carter, Das, Garnello, & Charboneau, 1984; Heiny, Houston, & Cooney, 1985; Houston, Intarapanich, Thomas, & Heiny, 1989; Intarapanich, 1988). A large number of studies have combined male and female faculty members into one regression model using dummy variables for the discriminatory factors (e.g. Braskamp & Johnson, 1977). Academic yearly salary is usually the criterion variable. A formerly discriminatory set of variables, market or discipline factors, are now accepted as justifiable reasons for salary differences (Gordon, Morton & Braden, 1976). Age, which some view as a proxy variable for experience, is considered by others as discriminatory (Heiny, et al., 1985; Snyder, Hyer & McLaughlin, 1993). Some consider rank a proxy variable for productivity since it correlates well with scholarly activity, research, and publications (Tennessee Higher Education Commission, 1979). Others think it is a "tainted" variable because discrimination in promotion could also occur along with discrimination in salary. For this purpose, Heiny, et al. (1985) have used canonical analysis and discriminant analysis to test the relationship between rank and sex and, also, to see if age might be a discriminatory variable. A related method for discerning relationships between sets of variables is recommended by Cohen (1993). Set correlation determines the proportion of generalized variance of one set of variables (dependent) accounted for by a second set of variables (independent). Besidesdiscipline factors and rank (if not related to sex), other acceptable explanatory variables are degree, tenure status, and experience.

Most researchers prefer to have a "fixed" model built with their preselected variables as Moore (1992) suggests. Computer selection techniques (stepwise regression, forward selection, backward elimination and all-possible-regressions) can be used to produce models that include only statistically significant variables. Baldus and Cole (1987) recommend deleting variables from the model to solve multicollinearity problems, but using fewer variables may mean a decrease in the predictive power. The number of observations available for a study also affects the number of independent variables to include in the regression equation. Crosswhite (1972) has shown that three subjects per variable is sufficient for

samples from populations whose coefficient of multiple determination is as low as 0.20.

Simpson and Rosenthal (1982) have suggested some standards that a final model should meet: a coefficient of multiple determination of 0.75 or more, a standard error of the estimate (SEE) less than 3000 For each institution separate equations containing all selected variables were developed. A variance inflation factor (VIF) greater than 10 and a condition number (CN) greater than 30 indicated moderate to severe collinearity. These equations were also subjected to various model selection techniques.

Three statistical techniques were used to determine the presence of "tainted" variables. Canonical analysis (CA) and set correlation were used to examine the relationship between the variables of gender and age and the nondiscriminatory variables. Structure coefficients of 0.30 or more signified an influence on the canonical variable. Set correlation measured the amount of variation in the set of variables, age and gender, that was explained by the other set of variables (nondiscriminatory). If this correlation was significant, both age and gender were tested separately. Discriminant analysis (DA) was used to determine possible misclassification of faculty members in both rank and tenure status.

Residual analyses were used to investigate the adherence of any prospective final model to the assumptions of multiple linear regression. These include linearity of the variables, normality of the error terms, and homoscedasticity.

Results

The data from each college was subjected to the procedures described in the methodology section. Descriptive statistics for each are given in Tables 3 and 4. and only statistically significant variables. However, Paetzold and Willborn (1994) have stated that an R^2 of 0.45 may be acceptable if the residual analysis confirms the applicability of the model (random residuals and absence of defects such as multicollinearity). For smaller institutions (N<100), an applicable model may be especially difficult to construct.

This research involved attempts to develop multiple linear regression equations to represent separate salary patterns for two small higher-education institutions . Records from the academic year 1992-93 for college A (N = 91) and the year 1993-94 for college B (N = 44) provided the information given in Tables 1 and 2. A new variable was developed for each college (average percent increase in salary per year).

As shown in Table 5 the initial model for college A including all selected variables exhibited multicollinearity.

Canonical, set correlation, and multiple discriminant analyses were conducted on this beginning model. The one significant canonical variate (CV) (Table 6) correlated highly with age and with the variables ESALA, PROFA, ASSTA, YRRKA, LONGA, TENA, TTA, and YRWTEN. The CV also had a structure coefficient of 0.3419 with gender. This result was consistent with the significant $R_{Y,X}^2$ of 0.727 (p = 0.000) from set correlation and the significance of the age variable (p=0.000). The p-value of the gender variable was 0.075.

In the multiple discriminant analysis for type of appointment, there were thirteen misclassifications, but only two of these were instances where the person should have been tenured and was not; one was a male and one was a female. Discriminant analysis on rank produced a different male and female in a lower rank than predicted.

Using the model selection techniques and dropping correlated variables, a "fixed" model was developed (ESALA, PROFA, ASSOCA, ASSTA, LONGA, TENA, GENDERA, BUSA, HUMA, EDUCA, MATHSCIA, HISTA, PSYCHA, PERF1A, and PERF2A). VIF for LONGA was still high (10.1), but to decrease it, ESALA would have to be dropped from the model. It acted as a suppressor variable so a large decrease in R^2 occurred, from 0.9236 to 0.7747 when ESALA was dropped, and the SEE increased from 2200.97 to 3753.32. This final model was checked for normality, linearity, and constant variance. The Shapiro-Wilk W was 0.9539 which had a p-value of 0.0100 indicating departure from normality. Also, entry salary exhibited a curvilinear relationship with salary (Figure 1), and the graph of the fitted values against the standardized residuals demonstrated an increase in variance as salaries increased. Using average percent increase in salary per year instead of entry salary gave a model that did not violate normality assumptions, but in the various different variable selection procedures, PCINCA was insignificant. Transforming salary (log salary, square root salary, inverse salary) did not solve the heteroscedasticity problem.

The original model for college B given in Table 7 had a very low adjusted R-square and a large SEE and exhibited multicollinearity. As with college A, the canonical correlation analysis of age and gender versus the other independent variables (Table 8) yielded one significant canonical variate (0.9160, p = 0.0006).

Table 1 College A Explanation of Variables

Variable	Description and/or Code
SALA	Academic yearly salary
ESALA	Entry salary
PROFA	1 (professor), 0 (else)
ASSOCA	1 (associate professor), 0 (else)
ASSTA	1 (assistant professor), 0 (else)
INSTA	1 (instructor), 0 (else)
YRRKA	Years in current rank
HPROFA	1 (hired professor), 0 (else)
HASSOCA	1 (hired assoc. professor), 0 (else)
HASSTA	1 (hired assistant professor), 0 (else)
HINSTA	1 (hired instructor), 0 (else)
AGEA	Chronological age
TTA	1 (tenure track, nontenured), 0 (else)
NTTA	1 (nontenure track), 0 (else)
TENA	1 (tenured), 0 (else)
YRWTENA	Years with tenure
LONGA	Length of service with institution
DOCA	1 (doctor's degree), 0 (else)
MAA	1 (master's degree), 0 (else)
GENDERA	1 (male), 0 (female)
BUSA	1 (business), 0 (else)
EDUCA	1 (education), 0 (else)
HISTA	1 (history), 0 (else)
HUMA	1 (humanities), 0 (else)
MATHSCIA	1 (mathematics or science), 0 (else)
PSYCHA	1 (psychology), 0 (else)
PVAA	1 (perform. and vis. arts), 0 (else)
PERF1A	1 (rating of 1), 0 (else)
PERF2A	1 (rating of 2), 0 (else)
PERF3A	1 (rating of 3), 0 (else)

Table 2 College B Explanation of Variables

Variable	Description and/or Code
SALB	Academic yearly salary
ESALB	Entry salary
PROFB	1 (professor), 0 (else)
ASSOCB	1 (associate professor), 0 (else)
ASSTB	1 (assistant professor), 0 (else)
INSTB	1 (instructor), 0 (else)
YRRKB	Years in current rank
HASSOCB	1 (hired associate professor), 0 (else)
HASSTB	1 (hired assistant professor), 0 (else)
HINSTB	1 (hired instructor), 0 (else)
AGEB	Chronological age
TENB	1 (tenured), 0 (else)
NTENB	1 (nontenured), 0 (else)
YRWTENB	Years with tenure
LONGB	Length of service
DOCB	1 (doctor's degree), 0 (else)
MAB	(master's degree), 0 (else)
GENDERB	1 (male), 0 (female)
BUSB	1 (business), 0 (else)
EDUCB	1 (education), 0 (else)
HISTB	1 (history), 0 (else)
HUMB	1 (humanities), 0 (else)
MATHCSB	1 (math or comp. science), 0 (else)
PSYCHB	1 (psychology), 0 (else)
PVAB	1 (perform. and vis. arts),0
(else)PRB	1 (park or recreation), 0 (else)
PHILB	1 (philosophy or religion), 0 (else)
SCIB	1 (sciences), 0 (else)

Table 3 College A Descriptive Statistics for Quantitative Variables

Females N = 30

	SALA	ESALA	AGEA	YRRKA	YRWTENA	LONGA
Mean	34270	22470	45.7	4.6	4.9	7.3
SD	6004	8575	7.9	5.5	8.1	7.8

Males N = 61

	SALA	ESALA	AGEA	YRRKA	YRWTENA	LONGA
Mean	37320	23120	45.7	6.1	5.9	8.7
SD	7660	8478	8.5	6.3	8.6	8.8

Table 4 College B Descriptive Statistics for Quantitative Variables

	SALB	ESALB	AGEB	YRRKB	YRWTENB	LONGB
Females N = 15						
Mean	27340	22740	46.6	3.7	1.1	4.9
SD	3464	6892	9.3	3.0	2.1	3.1
Males N = 29						
Mean	31380	18800	46.3	7.3	6.1	12.0
SD	4287	8530	8.5	8.4	7.8	9.5

Table 5 College A Multiple Linear Regression (All Variables)

Variable	*Coefficient*	*Std. Error*	*P*	*VIF*	*TOL*
Intercept	4112.12	4366.2	0.349		
ESALA	0.98	0.10	0.0000	11.4	0.09
PROFA	2984.97	1944.31	0.1294	13.7	0.07
ASSOCA	1472.70	1617.16	0.3657	8.6	0.12
ASSTA	908.80	1376.61	0.5114	7.6	0.13
HPROFA	-1237.80	2930.17	0.6741	1.6	0.62
HASSOCA	-1929.14	1373.22	0.1647	5.6	0.18
HASSTA	-995.01	966.28	0.3068	4.0	0.25
YRRKA	16.62	69.71	0.8123	3.1	0.33
LONGA	1336.58	151.25	0.0000	28.4	0.04
AGEA	2.39	45.86	0.6270	2.5	0.41
TENA	3791.16	3129.64	0.2300	42.4	0.02
TTA	1469.65	2992.81	0.6250	38.6	0.03
YRWTENA	-28.07	115.06	0.8080	16.1	0.06
DOCA	229.57	967.07	0.8131	3.3	0.30
GENDERA	449.18	633.12	0.4805	1.5	0.65
BUSA	2129.62	1135.67	0.0651	2.4	0.42
HUMA	536.73	1001.02	0.5936	2.0	0.50
EDUCA	96.35	932.36	0.9180	2.6	0.38
MATHSCIA	612.94	955.11	0.5232	2.6	0.38
HISTA	630.45	1097.99	0.5678	1.9	0.53
PSYCHA	-124.12	1143.38	0.9139	1.8	0.55
PERF1A	4022.70	1010.72	0.0002	2.3	0.43
PERF2A	1785.26	745.53	0.0194	2.1	0.48

R-Squared 0.9264 Adjusted R-Squared 0.9011
Standard Error of Estimate 2285.25

Table 6
Structure Coefficients for the Canonical Correlation Analysis of the Initial Model for College A

	V1	V2
AGEA	0.9411	-0.3381
GENDERA	0.3419	0.9397
	W1	**W2**
ESALA	-0.5674	0.3733
PROFA	0.8156	-0.0381
ASSOCA	-0.1811	0.0517
ASSTA	-0.5379	0.0146
HPROFA	0.1748	0.0535
HASSOCA	0.2025	0.2939
HASSTA	-0.1323	-0.1034
YRRKA	0.6124	-0.0909
LONGA	0.7459	0.2413
TENA	0.6150	0.0588
TTA	-0.6261	0.0087
YRWTEN	0.7547	-0.2910
DOCA	0.2612	0.0076
BUSA	0.0581	- 0.0839
HUMA	-0.0243	- 0.1298
EDUCA	-0.0885	0.0016
MATHSCIA	0.0233	0.3591
HISTA	0.1324	0.0795
PSYCHA	0.2267	-0.1797
PERF1A	-0.0156	-0.5539
PERF2A	0.0838	0.1328

Figure 1
Scatterplot of entry salary versus salary for College A

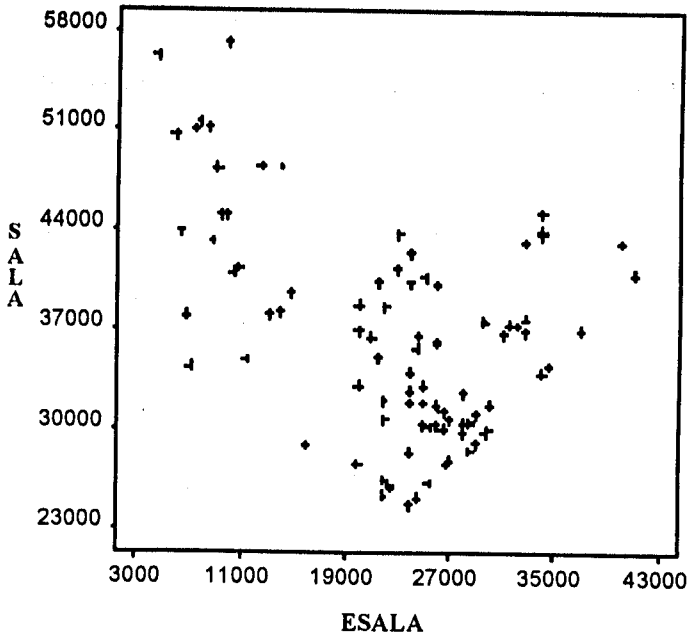


Table 7
College B Multiple Linear Regression (All Variables)

Variable	Coefficient	Std. Error	P	VIF	TOL
Intercept	18599.31	8458.83	0.0392		
ESALB	0.26	0.23	0.2652	13.1	0.08
PROFB	5051.59	4492.88	0.2736	11.4	0.09
ASSOCB	37.66	4037.89	0.9926	13.9	0.07
ASSTB	-1205.42	3468.59	0.7317	11.2	0.09
HASSOCB	5127.49	4219.32	0.2378	6.8	0.15
HASSTB	3402.35	2142.58	0.1272	3.9	0.26
YRRKB	73.28	246.51	0.7692	11.7	0.09
LONGB	329.33	426.54	0.4487	49.1	0.02
AGEB	-72.78	129.49	0.5800	4.7	0.21
TENB	705.26	1969.11	0.7238	3.6	0.27
YRWTENB	-207.87	347.93	0.5566	20.8	0.05
DOCB	1552.93	2622.36	0.5600	4.9	0.20
GENDERB	416.93	2056.65	0.8413	3.6	0.28
BUSB	5840.83	4238.15	0.1827	6.9	0.15
HUMB	863.25	2229.20	0.7025	1.9	0.53
EDUCB	2678.32	3058.70	0.3911	2.9	0.34
MATHCSB	4891.38	2762.71	0.0912	3.4	0.29
SCIB	1280.54	2482.55	0.6114	1.9	0.52
HISTB	-742.72	2965.45	0.8047	2.1	0.47
PSYCHB	-1448.78	2929.74	0.6261	2.1	0.48
PHILB	1843.74	2661.19	0.4960	1.7	0.59
PRB	-439.21	2855.68	0.8792	2.0	0.51

R-Squared 0.7113
Adjusted R-Squared 0.4088
Standard Error of Estimate 3406.63

Source	*DF*	*SS*	*MS*	*F*	*P*
Model	22	6.004E+08	2.729E+07	2.352	0.0274
Error	21	2.437E+08	1.161E+07		
Total	43	8.441E+08			

Table 8
Structure Coefficients for the Canonical Correlation Analysis of the Initial Model for College B

	V1	V2
AGEB	0.7880	0.6157
GENDERB	0.6045	-0.7966
	W1	W2.
ESALB	-0.4127	0.0054
PROFB	0.4642	-0.1452
ASSOCB	0.1754	-0.2066
ASSTB	-0.4070	0.2048
HASSOCB	0.4261	-0.0382
HASSTB	0.2522	0.0885
YRRKB	0.6249	0.1875
LONGB	0.7119	-0.0015
TENB	0.5190	-0.2042
YRWTENB	0.6554	0.0190
DOCB	0.2620	0.0118
BUSB	-0.4159	0.0612
HUMB	-0.1038	0.0227
EDUCB	0.1394	0.6199
MATHCSB	0.1681	-0.0938
SCIB	0.1472	0.0369
HISTB	-0.0637	0.2667
PSYCHB	0.0002	-0.0075
PHILB	-0.0731	-0.4161
PRB	-0.0364	-0.0433

Age had a structure coefficient of 0.7880 and the following variables had loadings of 0.30 or more: ESALB, PROFB, ASSTB, HASSOCB, YRRKB, LONGB, TENB, YRWTENB, and BUSB. However, gender also had a high correlation with this CV (0.6045). The $R_{Y,X}^2$ (0.919) from set correlation was significant ($p = 0.001$). Of the two Y variables, age was significant ($p = 0.009$) while the test for gender had a p-value of 0.071.

The multiple discriminant analysis for tenure status had two misclassifications, both males. There were three individuals (two males, one female) classified in ranks lower than they currently held and two males were classified into higher ranks.

For the model selection procedures, in all but backward elimination, the gender variable entered the model. However, in each case, its coefficient was insignificant after other variables were added. Gender had the highest correlation (0.4375) with salary of all the variables considered for the original model.

For this college, ESALB again acted as a suppressor variable, but LONGB also was a suppressor variable. They both had higher partial R-squares when they united with the variable PROFB. When all three were combined with gender in a stepwise regression, gender became insignificant or dropped out, suggesting that their presence may mask the significance of gender. Without ESALB in the model, LONGB did not exhibit a suppressor effect with PROFB. All models indicated that the variance was not constant, and models with ESALB violated the normality assumption.

Table 9 is a model in which ESALB was replaced with the average percent increase in salary per year (PCINCB). In all the variable selection procedures, PCINCB was significant. The "fixed" model (PCINCB, ASSOCB ASSTB, INSTB, HASSOCB HASSTB, GENDERB BUSB, HUMB, EDUCB, MATHCSB, SCIB, HISTB, PSYCHB, PHILB, and PRB) had an R^2 of 0.7775 (RSQ-adj = 0.6456) and a SEE of 2637.49. In the set correlation analysis ($R_{Y,X}^2 = 0.821$, $p = 0.002$), however, the other independent variables were shown not only to be related to age ($p = 0.015$), but also to gender ($p = 0.050$). The Shapiro-Wilk W was 0.9792 with a p-value of 0.7342, therefore, normality could be assumed. The residual plot still displayed variance that was not constant, though, and, as with college A, salary transformations did not provide any improvement.

Attempts were made to develop models for each of the institutions. Both initial models had severe collinearity. This was solved by removing variables that were intercorrelated with each other. Linearity and normality problems were also corrected. However, both sets of data demonstrated heteroscedasticity which was not remedied.

Table 9

College B Multiple Linear Regression (Percent Increase)

Variable	*Coefficient*	*Std. Error*	*P*	*VIF*	*TOL*
Intercept	18611.40	6220.02	0.0069		
PCINCB	38332.60	11846.50	0.0040	1.6	0.63
PROFB	1112.68	3735.21	0.7687	11.1	0.09
ASSOCB	-2512.13	3354.57	0.4622	13.5	0.07
ASSTB	-1713.51	2856.81	0.5551	10.7	0.09
HASSOCB	8378.17	3095.64	0.0132	5.2	0.19
HASSTB	3478.36	1750.60	0.0601	3.7	0.27
YRRKB	-95.15	212.17	0.6584	12.2	0.08
LONGB	26.97	292.86	0.9275	32.7	0.03
AGEB	25.32	108.73	0.8181	4.7	0.21
TENB	263.92	1668.17.	0.4571	3.7	0.27
YRWTENB	-5.62	285.18	0.9845	19.7	0.05
DOCB	1636.74	206.86	0.4665	4.9	0.20
GENDERB	2565.22	1588.85	0.1213	3.0	0.33
BUSB	10609.60	3065.65	0.0023	5.1	0.20
HUMB	1814.16	1850.65	0.3381	1.8	0.56
EDUCB	5500.35	2239.72	0.0229	2.2	0.45
MATHCSB	6456.48	2169.15	0.0072	3.0	0.33
SCIB	1552.36	2081.39	0.4640	1.9	0.53
HISTB	1315.44	2463.14	0.5989	2.1	0.48
PSYCHB	944.05	2450.90	0.7040	2.0	0.50
PHILB	276.30	2298.87	0.9055	1.8	0.56
PRB	2229.81	2216.00	0.3258	1.7	0.59

R-Squared 0.7953
Adjusted R-Squared 0.5809
Standard Error of Estimate 2868.32

Source	*DF*	*SS*	*MS*	*F*	*P*
Model	22	6.713E+08	3.051E+07	3.71	0.0019
Error	21	1.728E+08	8.227E+06		
Total	43	8.441E+08			

Conclusions

The purpose of this study was to try to develop multiple linear regression models for salary patterns from two small population (N < 100) higher education institutions. The initial R^2 and adjusted R^2 for college A were greater than 0.90 and the SEE was less than 3000 even though there was high multicollinearity. The results of canonical correlation indicated that older faculty were more likely to be tenured professors and have more years of service, more years in rank, and more years with tenure. They would also be less likely to be assistant professors and have been hired with high salaries. The multiple discriminant analysis did not detect any gender discrimination in type of appointment or in promotion. The p-value (0.075) for the gender variable in set correlation, however, gave some evidence of a possible relationship between gender and other independent variables.

It was determined from the various variable selection procedures that seven variables were statistically significant: entry salary, length of service, tenure status, business discipline, and the two performance variables. Gender entered only the forward selection model and was not significant at the 0.05 level. Entry salary was replaced by average percent increase in salary per year. This corrected the nonlinearity and nonnormality of models. Transformations of salary to log salary, square root salary and inverse salary did not correct for unequal variance in the error terms, however.

In the CA for college B, the one significant canonical variable had high positive correlations with age and gender. Older people were more likely to be tenured full professors, had been hired as associate professors, had been professors and had tenure longer, and had been at the institution longer. They also weren't as likely to have high entry salaries, be assistant professors or be in the business discipline. The high correlation for gender might mean that males also were more likely to exhibit these characteristics than females. Gender discrimination in tenure status or promotion was not signified in the DA. But, the fact that the gender variable had one of the highest correlations with salary and the circumstance that certain variables (entry salary, length of service, and professor) could mask this relationship signaled possible gender discrimination in salary.

Removing the entry salary variable gave a model that adhered to the normality assumption, but the heteroscedasticity was still present and the model had little predictive ability ($R^2 < 0.60$). Taking the entry salary values and using them to compute the average yearly percent salary increases resulted in a "fixed" model with R^2 greater than 0.70 and SEE less than three thousand. Since there was still a problem with variance that was not constant, however, no specific predictions could be made.

For both of these institutions, a problem that was not resolved was the unequal variation at different levels of salaries. This presented a prediction difficulty since residuals or standardized residuals could not be used to indicate that faculty members were being paid more or less than their equally qualified peers. Each of these colleges would be advised to use a case-by-case approach for determining gender discrimination in salary. College B might be especially concerned with this.

A suggestion for further research would be to try to find a way to weight salaries at different levels so that a model with homogeneous variance might be produced. Also, since the entry salary in both models enhanced the R^2 for each model (larger in college A), further study should be made concerning its relationship to other variables (i.e., longevity). Just because the data from these two colleges did not conform to appropriate multiple linear regression models for salary patterns does not mean that all higher education institutions with small faculty populations (N < 100) would have similar problems. They can be studied, individually, as these were to determine the suitability of this approach.

References

Baldus, D. C. & Cole, J. W. L. (1980). *Statistical proof of discrimination.* Colorado Springs, CO: Shepard's Inc.

Baldus, D. C. & Cole, J. W. L. (1987). *Statistical proof of discrimination* (Cumulative supplement). Colorado Springs, CO: Shepard's Inc.

Boyd, M. (1979). *Rank and salary differentials in the 1970s: A comparison of male and female full-time teachers in Canadian universities and colleges.* Ottawa, Ontario: Association of Universities and Colleges of Canada.

Braskamp, L. A. & Johnson, D. R. (1977, April). *The use of a parity-equity model to evaluate faculty salary policies.* Paper presented at the Annual Meeting of the American Educational Research Association, New York, NY.

Carter, R. D., Das, R. S., Garnello, A. H., & Charboneau, R. C. (1984). Multivariate alternatives to regression analysis in the evaluation of salary equity-parity. *Research in Higher Education, 20*(2), 167-179.

Cohen, J. (1993). Set correlation. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Statistical issues.* (pp. 165-198). Hillsdale, NJ: Lawrence Erlbaum Associates.

Crosswhite, C. E. (1972). Minimal subject-to-predictor ratios in multiple linear regression (Doctoral dissertation, University of Northern Colorado, 1972). *Dissertation Abstracts International, 33,* 3371B.

Finkelstein, M. O. & Levin, B. (1990). *Statistics for lawyers.* New York: Springer-Verlag.

Gordon, N. M., Morton, T. E., & Braden, I. C. (1976). Faculty salaries: Is there discrimination by sex, race, and discipline? *American Economic Review, 64,* 419-427.

Greenfield, E. (1977). From equal to equivalent pay: Salary discrimination in academia. *Journal of Law and Education, 6(*1), 41-62.

Heiny, R. L., Houston, S. R., & Cooney, J. B. (1985). Longitudinal analysis of salary discrimination in higher education. *Multiple Linear Regression Viewpoints, 14(*1), 1-38.

Hengstler, D. D., Muffo, J. A., & Hengstler, G. A. (1982, March). *Salary equity studies: The state of the art.* Paper presented at the Annual Meeting of the Association for the Study of Higher Education. Washington, DC.

Houston, S. R., Intarapanich, P., Thomas, A., & Heiny, R. L. (1989, April). *Salary discrimination in higher education: Selective multivariate analyses.* Paper presented at the American Educational Research Association Meeting. Boston, MA.

Intarapanich, P. (1988). Discrimination patterns and statistical procedures in faculty salaries (Doctoral dissertation, University of Northern Colorado, 1988). *Dissertation Abstracts International, 50,* 857A.

Moore, N. (1993). Faculty salary equity: Issues in regression model selection. *Research in Higher Education, 34*(1), 107-126.

Paetzold, R. L. & Willborn, S. L. (1994). *The statistics of discrimination,* Colorado Springs, CO: Shepard's/McGraw-Hill, Inc.

Simpson, W. A. & Rosenthal, W. H. (1982). The role of the institutional researcher in a sex discrimination suit. *Research in Higher Education, 16*(1), 3-26.

Snyder, J. K., Hyer, P. B., & McLaughlin, G. W. (1993, May). *Faculty salary equity: Issues and options.* Paper presented at the Annual Forum of the Association for Institutional Research, Chicago, IL.

Tennessee Higher Education Commission. (1979). *Tennessee higher education commission staff study re: house resolution no. 107.* Nashville,TN: Author. (ERIC Document Reproduction Service No. ED 177 986).

Practical Applications of Hierarchical Linear Models to District Evaluations

Gary W. Phillips, Ph.D
National Center for Education Statistics, U.S. Department of Education
Eugene P. Adcock, Ph.D
Prince George's County Public Schools, Maryland

This paper provides a practical application of hierarchical linear models (HLM) in an evaluation of effective schools for a large school district (the Prince George's County School District in Maryland). The HLM model is used first to rank elementary schools on their effectiveness at improving student learning in reading and mathematics and is also used to evaluate which factors contribute to school effectiveness. Teacher training was found to be the largest factor at contributing to school effectiveness after controlling for school context variables (School poverty and percent minority). It was found that this approach not only provides a rigorous statistical procedure, but also was easy to communicate to education policy makers. Plans for future analysis are also include.

This paper presents a "value-added" study of the effectiveness of 119 Prince George's County elementary schools' reading and mathematics programs. As suggested by Bryk and Weisberg (1976), a "value-added" approach is based on the use of a growth model to estimate the amount of growth that would be expected for a group participating in an educational program or school if they did not participate but instead were in the "regular" or "comparison" program. The actual change of the participants is compared to the predicted change and the difference is the "value-added." This approach is particularly well suited for evaluating school effectiveness or program effectiveness in their natural setting.

Following recent school effect studies of McPherson, 1992, Sanders & Horn, 1994, and Raudenbush & Willms, 1995, our practical application of the value-added model focuses on the influences of school practices (vs school context) which provide instructional treatments that raise student academic achievement regardless of the level at which the students enter the educational venue. A value-added school Effectiveness Index (EI) for the county was obtained from a new analysis of the statewide 1994 Maryland School Performance Assessment Program (MSPAP) controlling for student family socio-economic status and school population's percent of student poverty (Adcock, 1995). Hierarchical linear modeling analysis results provided an EI value for ranking each school's performance as either "ineffective," "no value-added," or "effective." Additional analyses examined the impact of a variety of teacher, school, and student background variables on schools' instructional effectiveness.

Method

In order to achieve the goals of the evaluation it was decided that the best statistical methodology would be hierarchical linear modeling. Hierarchical linear model analyses are like statistical microscopes in that they allow researchers and policy makers to see relationships in the data unconfounded by other variables. For example, the study attempted to determine the effectiveness of schools at promoting student achievement with the effects of student SES and school poverty controlled. In addition HLM was used to assess which school variables contribute to school effectiveness. It should be noted that only extant data were used in the study. Plans are under way to expand the data base so that the influence of other variables (such as school resources and instructional practices) may be assessed.

Recent articles on HLM applications were helpful in conceptualizing and explaining the analysis to policy makers. For example, Raudenbush and Willms (1995) distinguished between *Type A* and *Type B* school effects. *Type A* effects are often the interest of parents and real estate agents, whereas *Type B* effects are of more interest to education policy makers and evaluators. In a *Type A* effect we consider a school effective when students do well "regardless of whether that school's effectiveness derives from the superb practice of its staff, from its favorable student

composition, or from the beneficial influence of the social and economic context of the community in which the school is located. But it would clearly be unfair to reward school staff purely on the basis of their *Type A* effects, given that the staff is only partly responsible for those effects" (p.310). The *Type B* effect is the effect of school practice on student learning unconfounded by school context variables. HLM models are ideally suited to estimate *Type B* effects because they provide an index of school practice variables (curriculum content, instructional practice, and school resources) after factoring out the influence of school context variables (student demographics, community characteristics). "The *Type B* effect is the effect school officials consider when evaluating the performance of those who work in the schools. A school with an unfavorable context could produce a large *Type B* effect through the effort and talent of its staff. The school would rightly earn the respect of school evaluators even though parents shopping for a large *Type A* effect might not want to choose that school" (p. 310).

Past Practice

Before proceeding with the HLM model it is instructive to review an approach that many other evaluators have used in the past. In order to rank the schools based on an index of *Type B* school effectiveness that is unconfounded with student and school poverty education researchers have often used an ordinary least-squares regression (OLS) equation which includes a school poverty measure. One example of this would be the following single level equation

$$Y_{ij} = \beta_0 + \beta_1(X_{.j} - X_{..}) + r_{ij}. \qquad (1)$$

In equation 1 β_0 represents the predicted level of student achievement when school ($X_{.j}$) poverty equals the grand mean, β_1 represents the effects of poverty for schools, and r_{ij} is the error term. The model is essentially an OLS regression model with schools as the unit of analysis.

When this equation is used then the usual measure of *Type B* effectiveness is the difference between the actual mean performance of the school and the predicted performance based on school poverty, i.e., $Y_{.j} - Y_{..} - \beta(X_{.j} - X_{..})$

Although these OLS estimates have often been used, they have several statistical problems in comparison to HLM models. In the first place, they are unbiased but less efficient. The HLM estimates are both unbiased and more efficient. This is accomplished in HLM through the Bayesian procedure which uses not only the data available within a school for the regression equation but also uses all available data from other schools. The regression equation for each school is a weighted composite based on the information available in that school and the information available in the entire data set. The relative weights from these two sources depend on the precision of the parameter estimates. As the sample size of the school increases the weight of the school information dominates the parameter estimate. A by-product of the HLM solution to providing more stable estimates in smaller schools, is the added benefit that HLM more clearly partitions the variance within- and between schools, disentangles hypothesis testing for student versus school effects, and provides a general, yet flexible, way of modeling even with large numbers of student and school variables.

HLM Model

Instead of using the above single level model, the Prince George's County Effectiveness School Evaluation used a two level HLM model to assess *Type B* effects.

Level I

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - X_{..}) + r_{ij}, \qquad (2)$$

where

Y_{ij} = MSPAP scale score for student I in school j,

β_{0j} = expected MSPAP score for a student whose value on X_{ij} is equal to the grand mean, $X_{..}$. β_{0j} is an adjusted mean for school j such that $\beta_{0j} = \mu_{y\,j} - \beta_{1j}(X_{ij} - X_{..})$,

β_{1j} = expected change in MSPAP scores for a unit change in SES (i.e., the expected difference between SES = 1 and SES = 0) in school j, and

r_{ij} = residual for student I in school j.

Level II

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(W_{1j} - W_{1.}) + \mu_{0j}, \qquad [3]$$

where

γ_{00} = expected MSPAP mean for a PG County schools for students whose $W_{1j} = W_{1.}$,

γ_{01} = the relationship between the expected

school mean achievement ($ß_{0j}$) and
percent poverty in the school (W_{1j}),

μ_{0j} = unique effect of school j on the average achievement after controlling for W_{1j}.

$$ß_{1j} = \gamma_{10}, \qquad (4)$$

where

γ_{10} = the fixed value of the slope ($ß_{1j}$) across all schools (pooled within-school regression coefficient).

The above HLM model is called a random-intercept model because the $ß_{0j}$ is assumed to vary randomly across the level II units (schools). However, in the model, the within-school slopes are assumed to be constant across schools.

An important by-product of the HLM model is that it can be used to derive an index of the *Type B* effectiveness of schools at raising academic achievement after controlling for relevant student and school level variables. Once the index of effectiveness is obtained then schools can be ranked according to this index. The current index only controls for student and school level poverty. However, as data become available, and policy makers decide which variables they would like to include, then the index can be refined in the future. The effectiveness index μ_{0j} used in the evaluation to date is derived by the following steps.

1. Substitute equations {3} and {4} into equation 2

$$Y_{ij} = \{\gamma_{00} + \gamma_{01} (W_{1j} - W_{1\,.}) + \mu_{0j}\} + \{\gamma_{10}\}(X_{ij} - X_{..}) + r_{ij},$$

$$\mu_{0j} = Y_{ij} - [\gamma_{00} + \gamma_{01} (W_{1j} - W_{1\,.}) + \gamma_{10} (X_{ij} - X_{..})].$$

2. Average over i within j,

$$\mu_{0j} = Y_{.j} - [\gamma_{00} + \gamma_{01} (W_{1j} - W_{1\,.}) + \gamma_{10} (X_{.j} - X_{..})]. \qquad (5)$$

The effectiveness index in equation 5 is a measure of the schools level of academic achievement after controlling for student background effects, $\gamma_{10} (X_{.j} - X_{..})$, and school context effects, $\gamma_{01} (W_{1j} - W_{1\,.})$. It can be interpreted as the difference between the school's actual mean performance and the school's expected mean performance (based on the achievement of other schools with similar levels of student and school poverty).

Results

The above index was calculated for all 119 elementary schools and is included in the full report. Schools that are more than one standard deviation above what is expected (based on their levels of poverty) are considered effective. Schools that are within one standard deviation are considered doing about as well as can be expected (no value-added). Schools that are one standard deviation below are considered ineffective and are not performing up to the levels of other schools with similar levels of poverty.

Figure 1 (mathematics) and figure 2 (reading) provide a graphic representation of the relationship between the schools actual observed average score and the schools level of poverty.

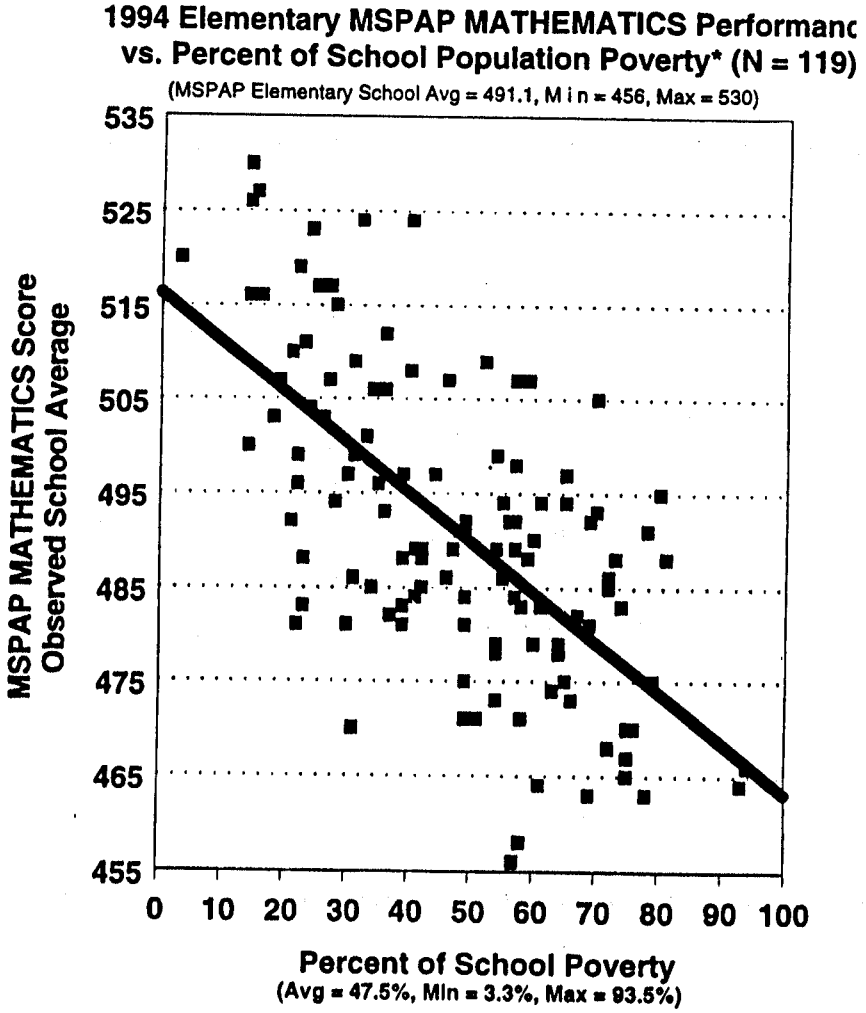
Figures 1 and 2 clearly show that there is a strong negative relationship (the correlation was -.70 for reading and -.64 for math) between the schools achievement and the population of poverty in the school. As the level of poverty goes down the school tends to achieve more. These graphs represent the type of *Type A* effects discussed above. Unfortunately, in figures 1 and 2 it is impossible to disentangle the effects of school practice from school context. An "evaluation" of schools would need to first control for school context variables. This is accomplished in figures 3 and 4 by controlling for student and school SES.

Figure 3 (mathematics) and figure 4 (reading) plot the effectiveness index against the schools level of poverty. These graphs are examples of *Type B* effects that school officials need in order to determine which schools have the most effective practice.

The data points above the upper boundary line in Figures 3 and 4 are those schools identified as effective while those below the lower boundary line are ineffective. It should be noted that at all levels of poverty, there are many schools that meet expectation (within one standard deviation), some that are effective (above one standard deviation) and some that are ineffective (below one standard deviation). It should be noted that the effectiveness index is not correlated with school poverty. This is why the effectiveness index as an accountability measure is an improvement over the mean MSPAP score. The average MSPAP score is highly correlated with school poverty and in fact 40% of the variance in school math performance can be attributed to school poverty as can 50% of the variance in reading. The important thing about figures 3 and 4 is that they provide a way of comparing schools with a more even

playing field. It shows that schools with similar levels of poverty have differing levels of student achievement. Some schools are not achieving well even though they have low levels of poverty and some are doing very well in spite of very high levels of poverty.

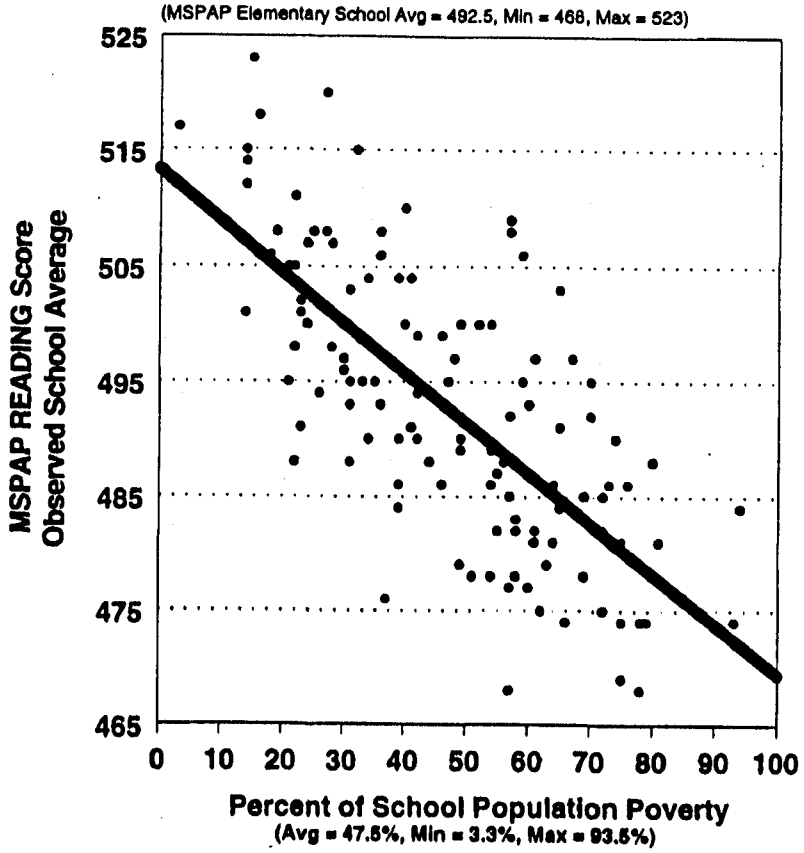
Figure 1



*** Correlation between school poverty and MSPAP Math = -.64**
*** School poverty calculated from Elementary grade students receiving F/R Meals and administered the MSPAP.**
Research, Evaluation & Accountability

Figure 2

1994 Elementary MSPAP READING Performance vs. Percent of School Population Poverty* (N = 119)

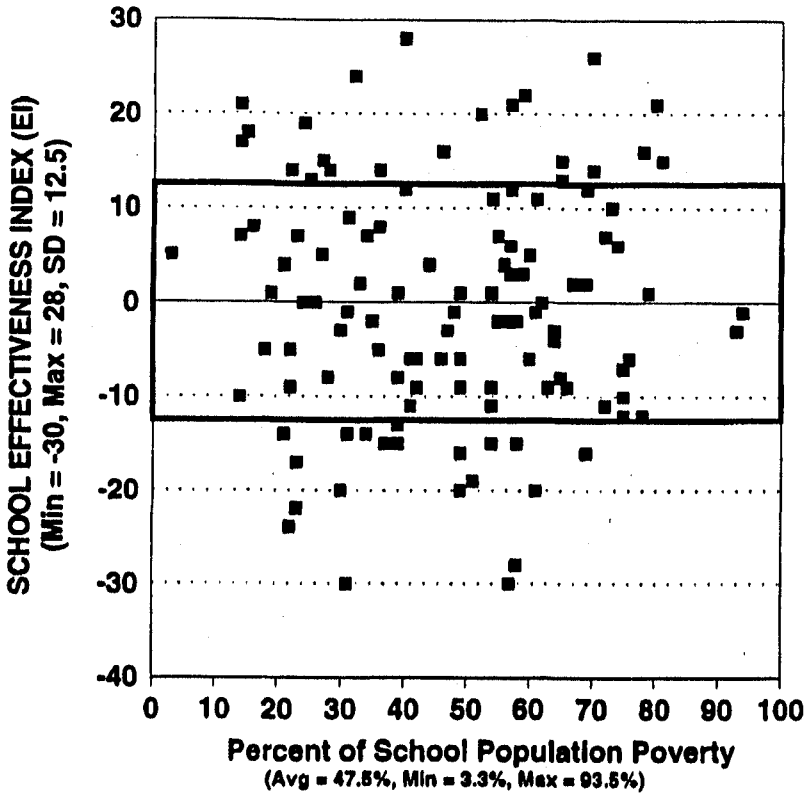


* Correlation between school poverty and MSPAP READING = -.70
* School poverty calculated from Elementary grade students receiving F/R Meals and administered the MSPAP.
Research, Evaluation & Accountability

Figure 3

MATHEMATICS EFFECTIVENESS INDEX (EI)
vs. School Population Poverty

1994 MSPAP Performance for 119 Schools

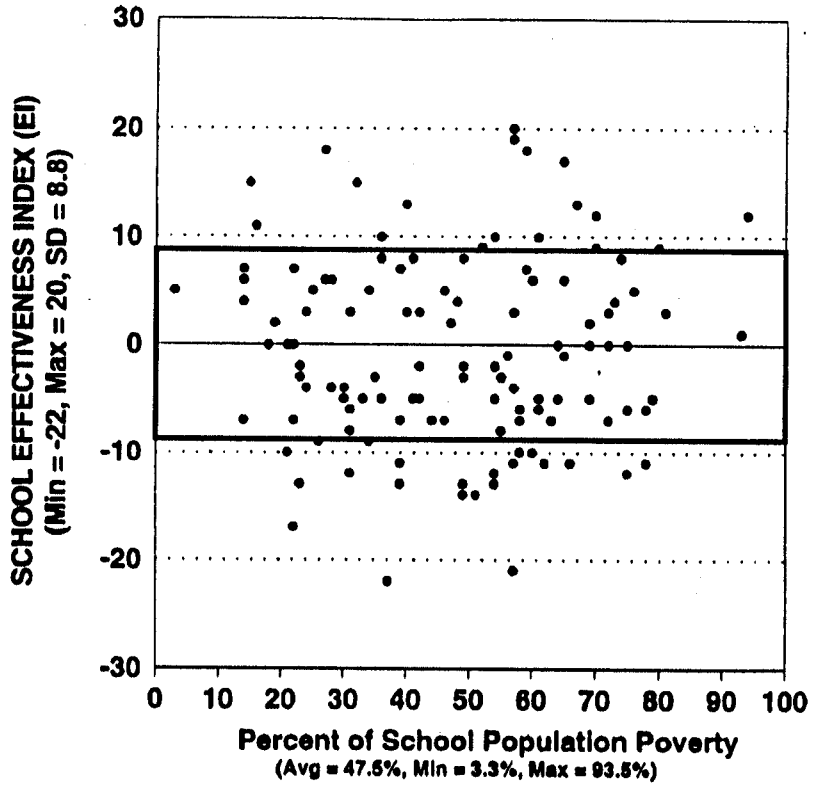


Note: Dots outside the middle zone indicate schools one or more standard deviations (+/- 12.5) from expected performance.**Correlation is ZERO**(r = 0.01, p = 0.94).
School poverty calculated from Elementary grade students receiving Free/Reduced Meals and administered the MSPAP.
Research, Evaluation & Accountability

Figure 4

READING EFFECTIVENESS INDEX(EI) vs. School Population Poverty

1994 MSPAP Performance for 119 Schools



Note: Dots outside the middle zone indicate schools one or more standard deviations (+/- 8.8) from expected performance. Correlation is ZERO (r = 0.02, p = 0.86).
School poverty calculated from Elementary grade students receiving Free/Reduced Meals and administered the MSPAP.
Research, Evaluation & Accountability

Additional Analyses

After ranking the schools based on the effectiveness index we also are interested in those variables that help to explain the rankings. The question we are attempting to answer is what factors help explain why some schools are more effective than others (i.e., a *Type B* effect based on school practice variables) after we have controlled for school context variables). This line of inquiry is only in its initial stages in the Prince George's County Evaluation. Additional data need to be collected that relate to additional school practices such as fiscal resources, teacher characteristics, instructional practices and curriculum offerings. However, as a first attempt at this analysis, extant school level data were used in which school poverty and percent minority are treated as school context variables and level of teacher training and Milliken status are treated as school practice variables. The HLM model that was fit to the data was as follows:

LEVEL I

$Y_{ij} = \beta_{0j} + \beta_{1j} (X_{ij} - X..) + r_{ij}$, where (6)

Y_{ij} = MSPAP score for student I in school j,

β_{0j} = expected MSPAP score for a student whose value on X_{ij} is equal to the grand mean, X.. . β_{0j} is an adjusted mean for group j such that $\beta_{0j} = \mu_{y\,j} - \beta_{1j} (X_{ij} - X..)$.

$ß_{1j}$ = expected change in MSPAP scores for a unit change in SES (i.e., the expected difference between SES = 1 and SES = 0) in school j, and

r_{ij} = residual for student I in school j.

LEVEL II

$$ß_{oj} = \gamma_{00} + \gamma_{01} (W_{1j} - W_{1.}) + \gamma_{02} (W_{2j} - W_{2.}) + \gamma_{03} (W_{3j} - W_{3.}) + \gamma_{04} (W_{4j} - W_{4.}) + \mu_{0j}, \text{ where} \quad (7)$$

γ_{00} = expected MSPAP mean for a non-Milliken school whose $W_{1j} = W1.$, $W_{2j} = W_{2.}$, $W_{3j} = W_{3.}$, $W_{4j} = 0$,

γ_{01} = the relationship between the expected school mean achievement ($ß_{0j}$) and percent poverty of the school (W_{1j}) after controlling for other school level variables,

γ_{02} = the relationship between the expected school mean achievement ($ß_{0j}$) and percent minority of the school (W_{2j}) after controlling for other school level variables,

γ_{03} = The relationship between expected school mean achievement ($ß_{0j}$) and levels of teacher training in the school (W_{3j}), after controlling for other school level variables,

γ_{04} = difference between Milliken and non-Milliken expected school mean achievement (W_{4j}) after controlling for other school level variables, and

μ_{oj} = unique effect of school j on the average achievement after controlling for W_{1j}, W_{2j}, W_{3j} and W_{4j}.

$$ß_{1j} = \gamma_{10} + \gamma_{11} (W_{1j} - W_{1.}) + \gamma_{12} (W_{2j} - W_{2.}) + \gamma_{13} (W_{3j} - W_{3.}) + \gamma_{04} (W_{4j} - W_{4.}), \text{ where} \quad (8)$$

γ_{10} = expected MSPAP slope for a non-Milliken school whose $W_{1j} = W1.$, $W_{2j} = W_{2.}$, $W_{3j} = W_{3.}$, $W_{4j} = 0$,

γ_{11} = the relationship between the expected school slope ($ß_{1j}$) and percent poverty of the school (W_{1j}) after controlling for other school level variables,

γ_{12} = the relationship between the expected school slope ($ß_{1j}$) and percent minority of the school (W_{2j}) after controlling for other school level variables,

γ_{13} = The relationship between expected school slope ($ß_{1j}$) and levels of teacher training in the school (W_{3j}), after controlling for other school level variables,

γ_{14} = difference between Milliken and non-Milliken expected school slopes (W_{4j}) after controlling for other school level variables.

To distinguish this more elaborate model (equations 6, 7 and 8) from the one used to rank the schools (equations 2, 3 and 4) we will refer to the earlier model as HLM_1 and the current model as HLM_2.

The results in Tables 1 and 2 are the primary findings from the fuller HLM_2 model. In each case an HLM analysis was conducted that included all available variables. Variables that did not show a significant relationship were deleted in the final model. The results for Table 2 were as follows: across all 119 schools as the percent of poverty increased 10% the mean math MSPAP score dropped 2.1 points; as the percent minority increased 10% the mean MSPAP score decreased 1.5 points; and, as the level of teacher training increased one level (e.g., from the bachelors to the bachelors plus 30 credit hours) the average MSPAP score increased 7 points. The level of teacher training was by far the variable with the strongest influence on the achievement of schools. It is also important to note that whether the school was a Milliken school was also a variable in the initial model. However, there was no significant difference in Milliken versus non-Milliken schools (after controlling for student SES and school poverty) so the variable was dropped in the final model. The results were similar, but less dramatic, for reading in Table 3.

In addition to assessing the influence of the effect of the above variables on average school achievement, the HLM_2 was also used to assess a question of equity. The issue here is the extent to which the schools achievement is really due to the SES of the student population. An index of this is captured by the level I $ß_1$ coefficients. The level I $ß_1$ coefficient represents the within -school relationship of student achievement to student SES. A large $ß_1$ indicates that there is a large relationship between achievement and SES within the school. A more desirable situation would be a small $ß_1$ which indicates that the

schools level of achievement is related to variables other than the SES of the student population. Tables 2 and 3 also contain these analyses. The results indicate: as the percent poverty increases 10% the β_1 decreases 1 point for math and 1.3 for reading; as the percent poverty increases 10% the β_1 drops 1 point for both math and reading; and as the average level of teacher training increases one level the β_1 increases by 4.7 points in reading. This last statistic is significant in that it means reading achievement is more related to the student's SES in the schools with the highest level of teacher training. This finding was not observed for math.

Table 1: Primary Findings for HLM_2 in Math

	Initial Model	Final Model
Model for Predicted School Means, β_0		
Intercept, _00	490.2 (1.2)	490.6 (1.1)
Percent Poverty, _01	-2.1 (0.6)*	-2.1 (0.6)*
Percent Minority, _02	-1.6 (0.6)*	-1.5 (0.6)*
Teacher Training, _03	7.0 (3.1)*	7.0 (3.2)*
Milliken Program, _04	2.0 (3.4)	
Model For SES Slope, β_1		
Intercept, _10	18.1 (0.9)*	18.2 (0.8)*
Percent Poverty, _11	-1.0 (0.5)**	-1.0 (0.5)*
Percent Minority, _12	-1.5 (0.65*	-1.7 (0.4)*
Teacher Training, _13	2.2 (2.3)	
Milliken Program, _14	0.6 (2.7)	

* There is at least a 95% chance that the true regression effect is not equal to zero.
** There is at least a 90% chance that the true regression effect is not equal to zero.

Table 2: Primary Findings for HLM_2 in Reading

	Initial Model	Final Model
Model for Predicted School Means, β_0		
Intercept, _00	492.5 (0.8)	492.2 (0.8)
Percent Poverty, _01	-1.8 (0.4)*	-1.8 (0.4)*
Percent Minority, _02	-1.0 (0.4)*	-1.1 (0.4)*
Teacher Training, _03	4.8 (2.1)*	4.9 (2.1)*
Milliken Program, _04	-1.4 (2.4)	
Model For SES Slope, β_1		
Intercept, _10	16.5 (0.9)*	16.4 (0.8)*
Percent Poverty, _11	-1.2 (0.5)*	-1.3 (0.5)*
Percent Minority, _12	-1.0 (0.5)**	-1.0 (0.5)*
Teacher Training, _13	4.6 (2.4)*	4.7 (2.4)*
Milliken Program, _14	-0.7 (2.8)	

* There is at least a 95% chance that the true regression effect is not equal to zero.
** There is at least a 90% chance that the true regression effect is not equal to zero.

Future Plans

The results described in this paper are based on an initial effort to evaluate the effectiveness of schools with a limited number of variables. Future plans include 1) increasing the grade levels to include both elementary and middle school, 2) including measures of science in addition to math and reading, 3) using both SES and percent minority as context variables, and 4) extending the number of school practice variables to include teacher training, teacher experience, fiscal resources available in the school, and educational effort from the student, teacher and the parents. In addition, more distant plans call for the use of a three-level HLM model in which change over time is modeled at the first level, student differences at the second level and school effects at the third level.

References

Adcock, E. P. (1995). *Value-Added Effective Schools Study for Elementary Schools: 1994 Maryland School Performance Assessment Program Results*, Research, Evaluation & Accountability, Prince George's County Public Schools, MD.

Bryk, A. S., & Weisberg, H. I. Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics,* 1976, 1, 127-155.

McPherson, A. F. (1992). *Measuring added value in schools* (NCE Briefing No. 1). London: National Commission of Education.

Raudenbush, S. W., & Willms, J.D. (1995) The estimation of School Effects, *Journal of Educational and Behavioral Statistics*, Winter 1995, Vol. 20, No. 4, pp. 307-335.

Sanders, W. L., & Horn, S. *The Tennessee Value-Added Assessment System (TVASS): mixed Model Methodology in Educational Assessment,* 1994.

MINUTES

OF THE
ANNUAL MEETING
OF THE
MULTIPLE LINEAR REGRESSION: GENERAL LINEAR MODEL / SIG
(New York, NY)

APRIL 11, 1996

Professor Isadore Newman (University of Akron), SIG Chair, opened the business meeting. The first order of business was the call for nominations of Chair-elect and two replacement Executive Board/Editorial Board members. Executive Secretary, Steve Spaner (University of Missouri - St. Louis), explained that the MLR:GLM/SIG election procedures call for the election to be held by mail ballot and the business meeting to be a nominating meeting only. It was moved and passed by the members attending to suspend the election by mail ballot rule and to hold the election at the business meeting. Nominations for chair-elect were Professor Dimiter Dimitrov (Kent State University) and Professor Jeffrey Kromrey (University of South Florida). Jeff Kromrey (University of South Florida) was elected Chair-elect for 1997. His term of office will begin following the 1997 business meeting. The nominated Executive Board/Editorial Board replacements were Professors Bruce Rogers (University of Northern Iowa), Dennis Leitner (Southern Illinois University-Carbondale), Jeffrey Hecht (Illinois State University-Normal), and Janet Sheehan (Northern Illinois University-Dekalb). Professors Dennis Leitner is renewed and Jeffrey Hecht replace Board member Susan Tracz (California State University-Fresno) and assume the four year terms from 1996 - 2000.

Chair Newman called upon Steve Spaner to give the treasurers report and the membership update. Spaner reported that the SIG treasury was $1917.93 on 4-1-95, the SIG account has earned $40.16 interest over the year and received $628.00 in member dues for a total assets of $2586.09. The SIG has incurred expenses of $297.08 since 4-1-95 leaving the SIG with a $2289.01 balance on 4-1-96. Spaner reported that the current paid membership was down from 1995. Spaner attributed the decline to the reduced number of issues of and irregular schedule for the Multiple Linear Regression Viewpoints (MLRV), the MLR:GLM/SIG's journal. Journal editor John Pohlmann urged members to submit articles and comments for consideration in MLRV. It was suggested, once again, that persons making presentations under the MLR:GLM/SIG sponsorship at the AERA conference should at least be invited to submit their papers to the MLRV.

Respectfully submitted,

Steven D. Spaner,
Executive Secretary

SPECIAL NOTICE

TO: LIBRARIES AND INSTITUTIONS (and MLR:GLM/SIG members)

RE: VOLUMES 18 - 23 (1991 - 96) of Multiple Linear Regression Viewpoints

The EBSCO and FAXON subscription services have been notified in each of the years listed above that the MLR Viewpoints has reduced its publication frequency to "occasional." While we strive to put out two issues per year (i.e., two issues per volume), for the past six years (six volumes) we have had insufficient submissions to make a second volume economical. We still hold to our goal of two issues a year, but do not guarantee two issues per year and do not honor claims for a second issue (i.e., the succeeding years' issue) in years when no second issue was published. We hope this clears up a number of outstanding claims notices. We thank you for your support of and interest in our journal and our Special Interest Group.

Sincerely,

John Pohlmann, PhD
Editor, MLR Viewpoints
Department of Educational Psychology
Southern Illinois University-Carbondale
Carbondale, IL 62901
e-mail: johnp@siu.edu

Steven Spaner, PhD
MLR:GLM/SIG Executive Secretary
Department of Behavioral Studies
University of Missouri-St. Louis
St. Louis, MO 63121-4499
e-mail: sspaner@umslvma.umsl.edu

(Secretary's note: 1997 membership payment is due at the beginning of the 1997 calendar year. If the first line of your mailing label ends in 96, you now owe for the 1997 MLR:GLM/ SIG membership year. If your mailing label has **95 or earlier** at the end of the first line, **you are unpaid** for the past 1996 MLR:GLM/SIG membership year as well as owe 1997 MLR:GLM/SIG membership dues)

Notice: 26Feb97

Information for Contributors

Multiple Linear Regression Viewpoints (MLRV), a publication sponsored by the American Educational Research Association's Special Interest Group on Multiple Linear Regression (SIG/MLR), is published once or twice per year to facilitate communication among professionals who focus their investigations on the theory, application, or teaching of multiple linear regression models or their extensions. Also, the journal accepts news items of interest to members of the SIG/MLR.

All manuscripts submitted to *MLRV* should conform to the language, style, and format of the *Publication Manual of the American Psychological Association* (4th ed., 1994), available from Order Department, American Psychological Association, P. O. Box 2710, Hyattsville, MD 20784. Three copies of the manuscript, all double spaced (including equations, footnotes, quotes, and references) and accompanied by an abstract of 100 words or less, should be submitted to the editor at the address listed below. Mathematical symbols and Greek letters should be precise and clear and should leave no question as to interpretation. All figures must be camera ready. Manuscripts that do not conform to the above specifications may be returned to the author for style changes before the review process will begin. A submitted manuscript will receive a blind review from at least two members of the editorial board (except occasional invited contributions, letters to the editor, editorials, or news items). Any author identifying information should appear on the title page only. Efforts will be made to keep the review process to a maximum of eight weeks. The final version of an accepted manuscript should be submitted on a 3.5" disk, preferably in Apple Macintosh Microsoft Word, Version 5.1, although other formats might be acceptable. The editor reserves the right to make minor changes to an accepted manuscript in order to facilitate a clear and coherent publication.

Potential authors are encouraged to contact the editor to discuss ideas for contributions or to informally determine whether manuscripts might be appropriate for publication in MLRV. The editor also welcomes suggestions for debates, theme issues, other innovative presentation formats, and general inquiries about the journal. SIG/MLR news items should be sent to the editor as soon as they become available.

Manuscripts and other correspondents with the editor should be addressed to:

John T. Pohlmann, Editor, *MLRV*
Department of Educational Psychology and Special Education
Southern Illinois Universtiy at Carbondale
Carbondale, Illinois 62901-4618

phone: (618) 536-7763
fax: (618) 453-7110
internet: JOHNP @ SIU.EDU

University of Missouri St. Louis

School of Education

Department of Behavioral Studies

8001 Natural Bridge Road
St. Louis, Missouri 63121-4499
Telephone: 314-516-5782

Postage paid at the University of Missouri at St. Louis, St Louis, MO 63121-4499. POSTMASTER Send address changes to Steven Spaner, MLR:GLM/SIG Executive Secretary, Department of Behavioral Studies, University of Missouri - St. Louis, 8001 Natural Bridge Road, St Louis, MO 63121-4499.

111 ROGE 98
ROGERS, BRUCE G.
DEPT OF ED PSYCH
UNIV OF NO IOWA
CEDAR FALLS IOWA 50614-0607