CLASSIFICATION BASED ON HIERARCHICAL LINEAR MODELS: THE NEED FOR INCORPORATION OF SOCIAL CONTEXTS IN CLASSIFICATION ANALYSIS

By

BRANDON K. VAUGHN* QIU WANG**

* Assistant Professor, The University of Texas at Austin.
** Doctoral Candidate in the Program of Measurement and Quantitative Methods, Michigan State University.

ABSTRACT

Many areas in educational and psychological research involve the use of classification statistical analysis. For example, school districts might be interested in attaining variables that provide optimal prediction of school dropouts. In psychology, a researcher might be interested in the classification of a subject into a particular psychological construct. The purpose of this study was to investigate alternative procedures to classification other than the use of discriminant and logistic regression analysis. A classification rule utilizing Hierarchical Linear Modeling (HLM) was derived and examined, with a following example which will show the benefit for using such an approach by comparing the hit rates to those of a logistic regression analysis. Specifically, a real data set on retention of Thailand students (7516 students in 356 schools) was investigated using a reduced logistic regression, full logistic regression, and multilevel model. The results show that a multilevel approach increases the level of correct classification. Suggestions for practical use are considered.

Keywords: Classification, Hierarchical Linear Models, Multilevel Models.

INTRODUCTION

Purpose

Many areas in educational and psychological research involve the use of classification statistical analysis. For example, school districts might be interested in attaining variables that provide optimal prediction of school dropouts. In psychology, a researcher might be interested in the classification of a subject into a particular psychological construct. The purpose of this study is to investigate alternative procedures to classification other than the use of discriminant and logistic regression analysis. A classification rule utilizing hierarchical linear modeling (HLM) will be derived and examined, with a following example which will show the benefit for using such an approach by comparing the hit rates to those of a logistic regression analysis.

Theoretical Framework and Educational Importance

The problem of classifying an observation arises in many areas of educational practice. Multivariate discriminant analysis is a commonly used procedure, as it is logistic regression. However, each procedure in their traditional form does not consider nested or repeated measures type data (Hair, Anderson, Tatham, & Black, 1988; Tabachnick & Fidell, 2001). Often, educational studies deal with nested data (students nested in teachers, teachers nested in schools, and so on).

Traditional classification rules applied to school related studies often view school characteristics as being uniformly applied to every student within a school and between schools (Raudenbush, 1988). However, this is typically not the case. In a typical school environment, various elements such as resources, qualifications, and time can vary from teacher to teacher, and school to school (Bidwell & Kasarda, 1980). Furthermore, each student might respond to learning or treatments differently. This would seem to imply that since school data tends to be multi-level in nature (Barr & Dreeben, 1983), then any applied statistical classification technique should also incorporate a multi-level approach. However, traditional classification rules do not incorporate multi-level data (Hair et al., 1988). In fact, most classifications based on multi-level data will tend to ignore one level of the multi-level dataset (typically level-2) and just classify

based on student level characteristics (Raudenbush, 1988). Other studies might never consider the use of multi-level data for classification, and only collect student-level data to begin with.

The benefit of such a multi-level approach to classification would not just benefit educational research. Other research areas might also benefit from a classification rule which incorporates nested data, such a medical or business research. Any situation in which multi-level data is expected might exhibit better hit rates if a classification rule is developed incorporating the multi-level perspective.

This paper will consider the issue of classification using a hierarchical linear modeling method. The use of this rule follows closely to the logic presented by Raudenbush and Bryk (2002) in their discussion of Bernoulli HGLM, yet the authors did not consider using their model for classification purposes. Raudenbush and Bryk's approach was primarily for understanding the effects of the different independent variables in the model and not classification. This research will present a formal rule for classification and hit-rates based on an HLM approach, thus focusing the use of the HLM model to one of individual prediction.

A review of literature has shown no other studies which have considered classification using an HLM approach. Although studies have been found where multi-level models were used with dichotomous outcomes (particularly medical studies), no studies were found that viewed the procedure as classification with a resulting hit-rate. There is a need for thorough research into the use of HLM for classification purposes, especially in comparison to traditional classification methods. Since most educational studies involving students often could be viewed from a multi-level perspective, a multi-level approach to classification might provide tighter hit-rates than traditional means.

A Review of Logistic Regression for Classification

This study considers the classification of an observation into one of two populations. Thus, the dependent variable is a dichotomous outcome of group membership. Logistic regression models a nonlinear probabilistic function of this dichotomous outcome (Neter, Wasserman, & Kutner, 1989). For the dichotomous outcome (Y_i = 0, 1) of an observation i with continuous independent variables

(X:X_{1i},X_{2i}, …, X_{ni}), X the probability of group membership (e.g., $Y_i = 1$) can be expressed as:

$$Y_i = \frac{e^{(\beta' \mathbf{X})}}{1 + e^{(\beta' \mathbf{X})}} \qquad (1)$$

For $\beta' \mathbf{X} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni}$. Dichotomous and categorical variables can also be included in this model with the use of dummy or indicator variables. The value of Y_i would be interpreted as the probability of an observation *i* belonging to Group 2. Another common expression of this model is that of a logit function which indicates the log odds ratio:

$$\eta_i = \ln\left[\frac{Y_i}{1-Y_i}\right] \beta' \mathbf{X} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} \qquad (2)$$

The non-linear nature of this model requires maximum likelihood estimation of parameters (Hair et al., 1988).

Once the parameters have been estimated, the model can be used to make predictions for new observations. The prediction of group membership for two groups, given observations X, is as follows: classify the observation into Group 2 if the predicted probability is larger than a specified value; otherwise, classify the observation into Group 1. In most applications of classification with logistic regression, the predicted probability is set at 0.5. This is particularly common when the two groups are approximately equal in regard to their population proportions (Fan & Wang, 1999). However, when the prior probabilities for the groups are known, a common procedure is to incorporate these priors as the cutoff point in the prediction rule. The cost of misclassification is also considered in the construction of the classification rule (Johnson & Wichern, 1988). For this research, the cost of misclassification is considered to be equal and thus not considered.

Once the classification rule has been established, the accuracy of the classifications can be considered (Tate, 1998). This is typically presented with a table indicating the number of correct and incorrect classifications from

applying the classification rule to the sample from which it was derived. Since the classification rule is known to perform consistently better on the data from which it is derived, a separate procedure called "cross-validation" is often suggested (Tabachnick & Fidell, 2001). A cross validation procedure uses the classification rule on a new sample of cases. Since having a new sample is often a practical limitation, a traditional cross-validation technique involves splitting the sample into two parts, one part to derive estimate the coefficients for the classification model, and the other part to validate the classification rates. Cross-validation techniques will not be considered in this research. Another classification check requires the use of jackknifing techniques. The estimation of the coefficients can be biased if used to assign a case from which the coefficients were derived. To avoid this, a jackknife procedure in which the classification model is derived without the case, and then used to classify the case, can be utilized. Jackknifing techniques will also not be considered in this research.

The Use of HLM for Classification

The HLM procedure that will be utilized for classification purposes is a Hierarchical Generalized Linear Model (HGLM). The use of this procedure follows closely to the logistic regression approach discussed earlier, and can be thought of as a multi-level extension of this technique. For the dichotomous outcome (Y_{ij} = 0, 1) of observation i with continuous independent variable X (X: X_{1ij}, X_{2ij},..., X_{nij}) nested within a level-2 grouping j with continuous independent variables W (W: W_{1j}, W_{2j},..., W_{nj}) the probability of group membership (e.g., Y_{ij} = 1) can be expressed as:

$$Y_{ij} = \frac{e^{(\beta'_j \mathbf{X}_j)}}{1 + e^{(\beta'_j \mathbf{X}_j)}} \qquad (3)$$

For $\beta'_j \mathbf{X}_j = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \ldots + \beta_{nj} X_{nij}$. As in the case of logistic regression, dichotomous and categorical variables can be included in either the level-1 or level-2 models. The value of Y_{ij} would be interpreted as the probability of an observation i in level-2 grouping j belonging to Group 2. Another common expression of this model is that of a logit function which indicates the log odds ratio:

$$\eta_{ij} = \ln\left[\frac{Y_{ij}}{1-Y_{ij}}\right] \beta'_j \mathbf{X}_j = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \ldots + \beta_{nj} X_{nij} \qquad (4)$$

This expression will be considered as the level-1 model. In this model, each level-2 grouping j can have a unique intercept and slope. The level-2 model can be expressed as

$$\beta_{Nj} = \gamma_{N0} + \sum_{s=1}^{S_N} \gamma_{Ns} W_{sj} + u_{Nj} \qquad (5)$$

For N=0,...n parameters. Each β_{Nj} can be modeled with a number (S_n) of level-2 predictors (W_{sj}), each with a possible random effect u_{nj}.

Once the parameters for both levels have been estimated, the model can be used to make predictions for new observations similar to logistic regression. The prediction of group membership for two groups, given level-1 observations X and level-2 observations W, would follow the same logic: classify the observation into Group 2 if the predicted probability is larger than a specified value; otherwise, classify the observation into Group 1. The decision for the cutoff point and cost of misclassification is also similar to the discussion for logistic regression. Classification accuracy is assessed by applying the multi-level prediction model to the sample. Alternative methods for determining classification rates could possibly be extended to this HLM model, but are not considered in this research.

Data source

The data that will be used to compare the multi-level classification to the logistic regression approach will be a data set first introduced by Raudenbush and Bryk (2002, p. 296). This data considers whether a child had to repeat a grade during primary years in Thailand. The survey consisted of 7,516 sixth grade students from 356 primary schools. The classification group consisted of two populations: "Repeat a grade", and "Not repeat a grade." The level-1 (student level) classification variables include socio-economic status, gender, type of dialect, breakfast status, and preprimary experience. The level-2 classification variables include school mean socio-economic status, size of the school, and availability of textbooks. A summary of the means and standard deviations for the variables in the data set are presented in Table 1.

	Variable Name	Mean	sd
Student-level variables			
Repetition	Y_{ij}	0.14	0.35
Socio-economic status	$(SES)_{ij}$	0.00	0.68
Gender	$(MALE)_{ij}$	0.51	0.50
Breakfast status	$(BREAKFAST)_{ij}$	0.84	0.36
Type of dialect	$(DIALECT)_{ij}$	0.48	0.50
Preprimary experience	$(PREPRIM)_{ij}$	0.50	0.50
School-level variables			
School mean SES	$(MEANSES)_j$	–0.01	0.44
School enrollment size	$(SIZE)_j$	0.00	0.85
Availability of textbooks	$(TEXTS)_j$	0.01	1.85

Table1. Descriptive Statistics For Grade-retention Data

Method

To illustrate the use of this rule, classification and hit-rates will be computed for the given data set after using HLM software to obtain estimation of the coefficients in the model. These results will be compared to a similar logistic regression analysis using all variables without regard to nesting characteristics in SPSS, and a logistic regression using only level-1 data. The hit-rates will be computed and compared. Since the data set involves some categorical explanatory variables, the linear discriminant analysis will not be used. Only the logistic regression analysis will be compared to the multi-level analysis. In order to compare the hit-rates between the two approaches, both logistic regression models will include the same variables as the HLM procedure, respectively.

For Raudenbush and Bryk's approach (2002), the authors hypothesized that lower repetition rates would be associated with the following level-1 variables: preprimary experience, socioeconomic status, gender, regular language status, and breakfast status. They also hypothesized that the following level-2 variables would repetition rates: school-mean SES, school size, and status of schools with textbooks for each student. This hypothesized model will be used for both the traditional logistic regression classification (using only level-1 data), and the conditional HLM model with both levels.

The logistic regression model for the level-1 dataset is

$$\eta_i = \beta_0 + \beta_1(SES)_i + \beta_2(MALE)_i + \beta_3(DIALECT)_i + \beta_4(BREAKFAST)_i + \beta_5(PREPRIM)_i \quad (6)$$

For this model, SES is grand-mean centered, and all other variables are dichotomized as dummy variables. Incorporating all level-1 and level-2 data in the logistic regression model yields

$$\eta_i = \beta_0 + \beta_1(SES)_i + \beta_2(MALE)_i + \beta_3(DIALECT)_i + \beta_4(BREAKFAST)_i + \beta_5(PREPRIM)_i + \beta_6(MEANSES)_i + \beta_7(SIZE)_i + \beta_8(TEXTS)_i \quad (7)$$

The level-1 conditional HLM model is

$$\eta_{ij} = \beta_{0j} + \beta_{1j}(SES)_{ij} + \beta_{2j}(MALE)_{ij} + \beta_{3j}(DIALECT)_{ij} + \beta_{4j}(BREAKFAST)_{ij} + \beta_{5j}(PREPRIM)_{ij} \quad (8)$$

For the level-2 model, Raudenbush and Bryk considered all coefficients as fixed, except for the constant coefficient. The level-2 model is represented by

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(MEANSES)_j + \gamma_{02}(SIZE)_j + \gamma_{03}(TEXTS)_j + u_{0j}, \quad \beta_{pj} = \gamma_{p0} \quad \text{for } p > 0. \quad (9)$$

The predicted probability is not se t at 0.5 for this dataset since the two groups are not approximately equal in regard to their population proportions. Research indicates that the prior probability of retention in Thailand is 0.15 (Raudenbush & Bhumirat, 1992). In addition, the sample estimate of population retention is close (0.14) to this value. Therefore, this value is used as the cutoff point in the prediction rule.

Results and conclusions

The data was first considered from a typical logistic regression classification analysis. First, only level-1 data were used in this analysis, which consisted of 7516 students. From this total, 1067 students needed to repeat a grade. Case analysis indicated no serious outliers in the data set. Although this particular logistic model is based on the level-1 model in the HLM approach to enable comparisons, tests of the model goodness of fit were considered. Both the Pearson chi-square teat and Hosmer-Lemeshow test indicated that the specified logistic model was correct. All other assumptions for logistic regression were considered and deemed satisfactory. The estimated logistic regression reduced model coefficients are shown in Table 2.

The overall relationship was significant based on a test with the full model of all five predictors versus a constant-only model ($\chi^2(.05; 5) = 127.06$, $p < 0.001$). The effects of two of the five hypothesized independent variables were significant at the 0.01 level: gender and pre-primer status. First, the null hypothesis that, on average, student logit is significantly different from zero is tested. The results were

Fixed effects	Coefficient	se	Wald Test (z)
Constant (predicted logit), β_0	1.73	0.10	305.89*
Gender-logit contrast, β_1	0.43	0.07	40.90*
Preprimer-logit contrast, β_2	0.61	0.07	80.74*
SES-logit slope, β_3	0.07	0.05	1.81
Dialect-logit contrast, β_4	0.04	0.07	0.28
Breakfast-logit contrast, β_5	0.06	0.09	0.42

* $p < .01$

Table 2. Results from the Reduced Logistic Regression Model

statistically significant (Wald z = 305.89, $p < 0.01$) indicating the average student logit (intercept) was necessary to describe the log odds ratio of repetition. The estimate mean intercept, $\hat{\beta}_0$, was −1.73 which indicates that the average logit for a reference student was −1.73, which translates into a probability of 0.15.

Next, the hypotheses of whether gender, pre-primer status, SES, type of dilect, and breakfast status are related to the logit of grade repetition were considered. The results for gender were statistically significant (Wald z = 40.90, $p < 0.01$), indicating that a significant difference exists between a males and females in regard to logit of repetition. The estimated logit contrast, $\hat{\beta}_1$, was 0.43 which indicates that males tended to have significantly higher logits than females after controlling for other variables in the model. The results for pre-primer status were also statistically significant (Wald z = 80.74, $p< 0.01$), indicating that a significant difference exists between a students with and without preprimary experience in regard to logit of repetition. The estimated logit contrast, $\hat{\beta}_2$ was –0.61 which indicates that students with preprimary experience tended to have significantly lower logits than students without preprimary experience after controlling for other variables in the model. All other effects were not significant. Since the main goal of this research is to compare classification rates between procedures, the other coefficients are not interpreted.

The classification results for the reduced logistic regression model are shown in Table 3. The overall classification based on the logistic regression model was not extraordinary. Considering the five predictor variables, the correct classification for non-repeating students was

		Predicted Outcome		
		No Repeat	Repeat	Total
Observed Outcome	No Repeat	4422 (68.6%)	2027 (31.4%)	6449
	Repeat	583 (54.6%)	484 (45.4%)	1067
		Overall correct assignment = 65.3%		

Table 3. Classification Table for Reduced Logistic Regression Model

68.6% and 45.4% for repeating students; the overall correct assignment was 65.3%. As seen by these percentages, cases tended to be overclassified into the largest group: non-repeat. For classifications in which the emphasis is on repeating students, this decision rule resulted in a 54.6% false positive prediction for those students who actually repeated their grade level.

Next, a logistic regression was run with the complete level-1 and level-2 data set, ignoring the nested nature of the data, as indicated in Equation 7. Once again, all assumptions for logistic regression were considered and deemed satisfactory. The estimated logistic regression full model coefficients are shown in Table 4.

The overall relationship was significant based on a test with the full model of all eight predictors versus a constant-only model ($\chi^2(.05; 8) = 138.52$, $p < 0.001$). The effects of four of the eight hypothesized independent variables were significant at the 0.05 level: gender, pre-primer status,

Fixed effects	Coefficient	se	Wald Test (z)
Constant (predicted logit), β_0	-1.76	0.10	309.40*
Gender-logit contrast, β_1	0.43	0.07	40.37*
Preprimer-logit contrast, β_2	-0.57	0.07	65.51*
SES-logit slope, β_3	-0.07	0.05	1.95
Dialect-logit contrast, β_4	0.04	0.07	0.32
Breakfast-logit contrast, β_5	0.06	0.09	0.40
MeanSES-logit slope, β_6	-0.25	0.09	7.00*
Size-logit slope, β_7	0.02	0.04	0.23
Textbook-logit slope, β_8	-0.04	0.02	4.41**

* $p < .01$, ** $p < .05$

Table 4. Results from the Full Logistic Regression Model

average school SES, and textbook availability. Note that textbook availability was not significant at the 0.01 level. First, the null hypothesis that, on average, student logit is significantly different from zero is tested. The results were statistically significant (Wald z = 309.40, p < 0.01) indicating the average student logit (intercept) was necessary to describe the log odds ratio of repetition. The estimate mean intercept, z = 309.40, p < 0.01) $\hat{\beta}_0$, was −1.76 which indicates that the average logit for a reference student was −1.76, which translates into a probability of 0.17.

Next, the hypotheses of whether gender, pre-primer status, SES, type of dilect, breakfast status, average school SES, school size, and textbook availability are related to the logit of grade repetition were considered. The results for gender were statistically significant (Wald z = 40.37, p < 0.01), indicating that a significant difference exists between a males and females in regard to logit of repetition. The estimated logit contrast, $\hat{\beta}_1$, was 0.43 which indicates that males tended to have significantly higher logits than females after controlling for other variables in the model. The results for pre-primer status were statistically significant (Wald z = 65.51, p < 0.01), indicating that a significant difference exists between a students with and without preprimary experience in regard to logit of repetition. The estimated logit contrast, $\hat{\beta}_2$, was –0.57 which indicates that students with preprimary experience tended to have significantly lower logits than students without preprimary experience after controlling for other variables in the model. The results for average school SES were statistically significant (z = 7.00, p < 0.01), indicating a significant relationship between school SES and logit of repetition. The estimated logit contrast, $\hat{\beta}_6$, was –0.25 which indicates Finally, the results for textbook availability were statistically significant (Wald z = 4.41, p < 0.05), indicating that a significant effect of textbook availability on logit of repetition exists. The estimated logit contrast, $\hat{\beta}_8$, was –0.04 which indicates a negative relationship between textbook availability and logit. All other effects were not significant.

The classification results for the full logistic regression model are shown in Table 5. The overall classification

		Predicted Outcome		
		No Repeat	Repeat	Total
Observed Outcome	No Repeat	4233	2216	6449
		(65.6%)	(34.4%)	
	Repeat	529	538	1067
		(49.6%)	(50.4%)	
		Overall correct assignment = 63.5%		

Table 5. Classification Table for Full Logistic Regression Model

based on this logistic regression model was also not extraordinary, yet better than the logistic regression model using only level-1 predictors for predicting repeating students. Considering the eight predictor variables, the correct classification for non-repeating students was 65.6% and 50.4% for repeating students; the overall correct assignment was 63.5%. As seen by these percentages, cases tended to be overclassified into the largest group: non-repeat. For classifications in which the emphasis is on repeating students, this decision rule resulted in a 49.6% false positive prediction for those students who actually repeated their grade level.

The HLM analysis was considered next. Assumptions for HLM were checked and deemed satisfactory to continue with the analysis. The HLM analysis first considers the conditional model, with no level-1 or level-2 predictors in the model. The probability of retention for a "typical" school was 0.097. This probability is based on a school-level random effect of 0 (u_{0j} = 0). Converting these unconditional estimates into a confidence interval, 95% of schools have repetition probabilities between (0.01, 0.59). That is, some schools have nearly no repeats, while other schools have over half of their students repeating a grade. The results of this unconditional analysis are not shown, and the reader is referred to Raudenbush and Byrk (2002) for more information of their results.

The results of the conditional hypothesized HLM analysis are presented in Table 6. For the fixed effects, the null hypothesis that, on average, student logit is significantly different from zero is tested. The results were statistically significant (t(352) = -15.68, p < 0.01) indicating the average student logit (intercept) was necessary to

describe the log odds ratio of repetition. The estimate mean intercept, $\hat{\gamma}_{00}$, was -2.03 which indicates that the average logit for a reference student was -2.03, which translates into a probability of 0.12. This value is lower than the estimated constant coefficient using logistic regression. Next, the hypotheses of whether gender, pre-primer status, SES, type of dilect, and breakfast status are related to the logit of grade repetition were considered. The results for gender were statistically significant (t(7507) $= 6.92$, $p < 0.01$), indicating that a significant difference exists between a males and females in regard to logit of repetition. The estimated logit contrast, $\hat{\gamma}_{10}$, was 0.51 which indicates that males tended to have significantly higher logits than females after controlling for other variables in the model. The results for pre-primer status were also statistically significant (t(7507)=-6.23,p<0.01), indicating that a significant difference exists between a students with and without preprimary experience in regard to logit of repetition. The estimated logit contrast, $\hat{\gamma}_{20}$, was –0.60 which indicates that students with preprimary experience tended to have significantly lower logits than students without preprimary experience after controlling for other variables in the model. All other level-1 effects were not significant. The results are parallel to logistic regression, although the estimation of coefficients yielded different results.

Level-2 effects are reported in Table 6. None of these effects were significant. That is, the logit of repetition was unrelated to school MEANSES, school size, and textbook availability. Formal tests of whether the τ is significantly greater than zero is also presented in Table 4. The estimated variance for school logit is $\hat{\tau}_{00} = 1.31$ and is statistically significant ($\chi^2(352) = 1428.71, p < 0.001$), This indicates that significant differences exist among the various school predicted logits of their students.

The classification results using the multi-level HLM model are shown in Table 7. The overall classification based on the HLM model was also not extraordinary. Considering the five level-1 and three level-2 predictor variables, the correct classification for non-repeating students was 77.7% and 34.6% for repeating students; the overall correct assignment was 71.5%. As seen by these

Fixed effects		Coefficient	se	t Ratio
Model for predicted logit				
INTERCEPT, γ_{00}		- 2.03	0.13	- 15.68*
MEANSED, γ_{01}		–0.26	0.19	–1.32
SIZE, γ_{02}		0.01	0.09	0.08
TEXTS, γ_{03}		–0.05	0.04	–1.20
Model for gender-logit contrast				
INTERCEPT, γ_{10}		0.51	0.07	6.92*
Model for preprimer-logit contrast				
INTERCEPT, γ_{20}		–0.60	0.10	–6.23*
Model for SES-logit slope				
INTERCEPT, γ_{30}		–0.09	0.05	–1.59
Model for dialect-logit contrast				
INTERCEPT, γ_{40}		0.04	0.07	0.53
Model for breakfast-logit contrast				
INTERCEPT, γ_{50}		–0.04	0.10	–0.40
Random effects	Variance	df	χ^2	p value
School logit, u_{0j}	1.31	352	1428.71	0.001
Reliability of OLS Regression Coefficient Estimate				
Logit of repeating, β_{0j}	0.68			

Note: df = 352 for all t tests for γ_{0j} ; otherwise, df = 7507.
* p < .01

Table 6. Results from the Conditional Model

		Predicted Outcome		
		No Repeat	Repeat	Total
Observed Outcome	No Repeat	5008 (77.7%)	1441 (22.3%)	6449
	Repeat	698 (65.4%)	369 (34.6%)	1067
		Overall correct assignment = 71.5%		

Table 7. Classification Table for HLM Model

percentages, cases tended to be overclassified into the largest group: non-repeat. For classifications in which the emphasis is on repeating students, this decision rule resulted in a 65.4% false positive prediction for those students who actually repeated their grade level.

Of particular interest to the research question is the comparison of classifications between logistic regression

and HLM. As seen by Tables 3, 5, and 7, there is a better overall hit-rate by incorporating a multi-level approach for classification purposes than by incorporating a reduced or full logistic regression model. Overall, there was at least a 5% change for the better by using the HLM approach to classify cases. Yet, this result seems to be due to the fact of better classifications for the non-repetition group. For non-repeating students, use of HLM resulted in nearly a 10% improvement in classification. For repeating students, the HLM approach actually resulted in higher misclassifications than either logistic approach. There were 34.6% correct classification of repeating students using HLM, while 45.4% and 50.4% correct classifications using logistic regression, respectively.

Practical Implications

A possible implication of the study is a classification procedure more efficient than traditional analysis procedures when there is the presence of multi-level data. The multi-level approach appears to be a better procedure to use in such cases, particularly for classification of larger groups, as shown by the results of this study. However, this study also indicated that HLM was not always better than a traditional logistic regression approach, particularly for classification of smaller groups. As mentioned previously, very little study has been conducted comparing a multi-level classification method to traditional one-level methods. Multi-level analyses are widely used in many circles (such as business and medical applications), and with more thorough comparisons with traditional approaches, a multi-level classification rule could become a viable alternative for many practitioners.

Limitations and Future Study

The limitations of the study include the following: (1) only one data set is being considered; (2) the only comparisons being made are that of the logistic regression analysis due to the presence of categorical explanatory variables; and (3) no simulation study is conducted to look at the efficiency of various analyses under differing assumptions and conditions. Replications of this study should be carried out with a variety of data sets and situations. Varying such factors as classification probability rule (for various priors), presence of assumption violations, and varying sample sizes for level-1 and level-2 data should be considered. The author is currently working on a simulation study to study these effects on classification hit-rates.

Alternative classification procedures (such as regression trees) should also be considered. The effects of classification into more than two populations would be an interesting extension of this idea, as well as the incorporation of repeated measures data. New types of classification measures should be derived for the HLM procedure if replication studies find it a viable classification technique. Techniques such as cross-validation and jackknifing should be developed for the HLM procedure. Current HLM software does not allow for such techniques to be applied to HLM results. Finally, there are many measures in one procedure that do not seem as widely used in the other procedure (e.g., goodness of fit tests). More work can be done to find the commonality between the procedures presented in this research.

One possible consideration is that single level models with fixed effects are known to give unbiased estimates. An HLM classification procedure might be more advantageous not for the estimation of fixed effects, but in the incorporation of random error. Thus, one possible benefit from a multi-level approach might be a stepwise model selection of a classification model, since this involves inference (which incorporates error estimates).

Summary and Conclusions

A logistic regression classification rule was compared to a hierarchical linear modeling approach. Hit rates for a popular data set were compared for both procedures. Based on these results, the HLM approach to classification was superior to the logistic regression for larger groups and overall, although both rules exhibited what some may consider as low hit-rates. However, this may be due to the model chosen, which was based on a previous research model. The use of HLM is most likely warranted only in situations in which multi-level data is likely. Yet, more studies are needed to assess the degree to which this

procedure might be more effective than traditional classification means.

References

[1]. Barr, R., & Dreeben, R. (1983). *How schools work.* Chicago: University of Chicago Press.

[2]. Bidwell, C., & Kasarda, J. (1980). Conceptualizing and measuring the effects of school and schooling. *American Journal of Education, 88*, 401-430.

[3]. Fan, X., & Wang, L. (1999). Comparing Logistic Regression with Linear Discriminant Analysis in Their Classification Accuracy. *Journal of Experimental Education*, *67*, 265-286.

[4]. Hair, J. F., Anderson, R.E., Tatham, R.L., & Black, W.C. (1988). *Multivariate Data Analysis*. Upper Saddle River, NJ: Prentice Hall.

[5]. Johnson, R. A., & Wichern, D. W. (1988). *Applied Multivariate Statistical Analysis* (2nd ed.). Englewood Cliffs, New Jersey: Prentice Hall.

[6]. Neter, J., Wasserman,W., & Kutner, M. (1989). *Applied linear regression models* (2nd ed.). Homewood, IL: Irwin Newport.

[7]. Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13*(2), 85-116.

[8]. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications, Inc.

[9]. Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics*. Needham Heights, MA: Allyn & Bacon.

[10]. Tate, R. (1998). *An Introduction to Modeling Outcomes in the Behavioral and Social Sciences*. Edina, Minnesota: Burgess International Group, Inc.

ABOUT THE AUTHORS

Brandon Vaughn's current research interests include: multi-level differential item functioning (DIF), Bayesian estimation procedures, creative uses of non-parametric classification procedures, and effective strategies in the teaching of statistics. He has developed several technological tools for teaching statistics, including free R tutorial videos and applets for conceptual understanding.

Qiu Wang is a Doctoral Candidate in the Program of Measurement and Quantitative Methods at Michigan State University. His research interests focus on large scale data analysis including propensity score matching, classification, structural equation modeling, and Bayesian hierarchical modeling.