

**Can Hierarchical Linear Modeling Be Used to Rank Schools:
A Simulation Study With Conditions under which
Hierarchical Linear Modeling is Applicable¹**

Dash Weerasinghe and Timothy Orsak

Dallas Public Schools
Dallas, TX 75204
March 8, 2006

¹ Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 1998.

Can Hierarchical Linear Modeling Be Used to Rank Schools: A Simulation Study With Conditions under which Hierarchical Linear Modeling is Applicable¹

Dash Weerasinghe and Timothy Orsak

Dallas Public Schools
Dallas, TX 75204
March 8, 2006

Hierarchical Linear modeling is becoming a widely used technique to model student achievement within and between schools. As such there have been no validation studies done on simulated data where hierarchical modeling is used to recover the characteristics of the original simulated data. This paper describes the results of a large scale simulation study where student data within schools are simulated with known student and school characteristics and hierarchical linear modeling is used to recover these characteristics. Included will be the limitations of HLM with respect to the minimum number of students needed per school to effectively recover the student characteristics, the smallest difference in effect size that can distinguish between two schools and the combination of the above two parameters that can lead to false results.

Introduction

The motivation for a simulation study of rankings of schools using student achievement data arose due to the varying number of students available with pre-test and post-test scores that could be used in an analysis. In most cases that we have encountered in our research at the final analysis stage, the students per classroom can be as low as five or as high as twenty-two. Another important factor to consider is the smallest Effect Size that distinguishes a school from the next school in the rankings. Does the smallest Effect Size that distinguishes two schools depend on N , the number of students per school? What is the confidence level that Hierarchical Linear Modeling will give a true ranking for data with a specific Effect Size and Class Size. This simulation study attempts to answer the above questions regarding school rankings.

The standard hierarchical approach to school ranking is to model within each school an outcome variable measuring student performance by the student's past performances and to model the resulting intercept and slopes for each school by school level characteristics

¹ Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA, April 1998.

to obtain the variation of the intercept around the overall mean intercept irrespective of the group. There are various statistical methods that could be used to carry out this analysis, and one of the most widely used methods is an iterative maximum likelihood process where within school intercepts and slopes are first calculated and these are then modeled among the schools. The resulting unbiased estimates are used to calculate the within school intercepts and slopes again. This process is repeated until successive iterations yield the same maximum likelihood function within a certain tolerance. This approach become feasible when the Estimation Maximization (EM) algorithm was developed in 1977 (Dempster, Laird & Rubin).

In prior research, the effect of sample size, N , has been analyzed with respect to the standard error in the covariance matrix (Snidgers & Bosker, 1993). In this study the authors give guidelines on the choice of analyzing many schools with few students or few schools with many students. In almost all other design studies, HLM has been used on large sample surveys and the question of “Do we have enough within-group data?” has not risen. This paper with its conclusions hopes to extend the use of HLM to a larger audience where sample sizes within each group are not “large”. De Leeuw and Kreft (1993) addressed many questions with multilevel models regarding computer programs employing different algorithms, interpretations of the coefficients and of most importance the fact that traditional techniques perform as well, or better, if there are large groups and small intraclass correlations. Thus, since we have alternatives for hierarchical models for large groups, the performance of HLM under small group sizes is of even more importance. Further when the data are balanced, restricted maximum likelihood estimates for hierarchical designs duplicate the classical ANOVA results, (Raudenbusch, 1995) and for large sample sizes ANOVA would be as satisfactory as HLM.

This paper describes the results obtained when the Class Size is varied from five students per classroom to twenty-three and when the Effect Size that distinguishes two adjacent schools in the rankings are varied as a percentage of the standard deviation of the student scores. When can hierarchical linear modeling be used to rank schools? Well, it cannot be used all of the time, but when the number of students per school and the distinguishing Effect Size between two schools meets a certain criterion, HLM can be used and used effectively to rank schools.

Methodology

The model used in this HLM study is as follows:

$$POST_{ij} = \beta_{0j} + \beta_{1j}(PRE_1)_{ij} + \beta_{2j}(PRE_2)_{ij} + r_{ij} \quad (1)$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (2)$$

$$\beta_{1j} = \gamma_{10} \quad (3)$$

$$\beta_{2j} = \gamma_{20} \quad (4)$$

where $i = 1, 2, \dots, N_j$, $j = 1, 2, \dots, 21$. There are twenty-one schools and N_j is the number of students in school j . PRE_1 and PRE_2 are the two pre-test scores used to predict the $POST$ test score for each student with r_{ij} the i^{th} student's residual within school j . The

parameter β_{0j} is the intercept for school j and is randomly varying, while β_{1j} and β_{2j} are the slopes and are assumed fixed. The Empirical Bayes residuals are obtained from the estimates of u_{0j} , \hat{u}_{0j} , with a reliability adjustment. These Residuals are normally distributed with mean zero.

The data have the following structure. The twenty-one schools have five classrooms of equal size n , resulting in $N = 5n$ students per school. The students' pre-test scores have a normal distribution with a mean of zero and a standard deviation of one. Post-test scores of students' in each classroom within a school have a normal distribution with a mean equal to the assigned classroom effect and a standard deviation of one. The classroom effects are accumulated to produce the school effect for each school, and these characterize the effectiveness of the school. This variation of classrooms within schools is used to simulate the actual makeup of schools, where no two classrooms are alike. These school effects characterize the effectiveness of the school and uniquely determines the ranking of each school within the 21 schools. The difference in school effects between any two consecutively ranked schools is defined as the Effect Size. Effect Size and Class Size are used as parameters in this study.

Once the Class Size and the Effect Size among schools are chosen, the students' pre-test and post-test data are simulated. The *PRE_1*, *PRE_2* and *POST* scores are randomly generated from a standardized normal distribution. The *POST* scores are then biased by each classrooms' effect and a HLM analysis carried out to produce the Empirical Bayes residuals for each school. These residuals are indicators of each school's performance, and are used to rank the schools. This ranking is compared with the true ranking of the schools to determine the effectiveness of Hierarchical Linear Modeling. This process is repeated for 100 simulations.

The above process is carried out for varying combinations of Class Sizes and Effect Sizes to determine the effect of these two parameters on the hierarchical linear model used to obtain school effectiveness. The Effect Size is varied from 0.001 to 0.010, where 0.001 is 0.1% of standard deviation of the test scores, and 0.010 is 1% of standard deviation of test scores. Thus an Effect Size of 0.005 signifies that two consecutive ranked schools differ by 0.005 in their assigned bias.

The classroom effects within schools are distributed as follows. In the case when Effect Size is 0.01, in School 1, the students' post-test scores in each classroom is biased with classroom effects of -0.104, -0.102, -0.100, -0.098 and -0.096, respectively, resulting in a mean school effect for School 1 of -0.100. A similar distribution is used for all classrooms within schools, with the mean bias for schools successively increasing by 0.01, which is the Effect Size chosen. Thus, the students' post-test scores in School 2 has a mean school effect of -0.090, school 3 a mean school effect of -0.080, school 11 a mean school effect of 0.0, and so on with School 21 having a mean school effect of +0.100. The post-test scores were biased in such a way that the mean bias of all students irrespective of classroom or school is zero. A correct HLM ranking of the schools will result in School 1 being the lowest ranked school and School 21 the highest ranked

school. The table below describes the distribution of the school effects and the classrooms effects within schools for an Effect Size of 0.01.

Table 1. The distribution of classroom effects within schools and schools effects for an Effect Size of 0.01.

School	Classroom	Classroom Effect	School Effect
1	A	-0.104	
	B	-0.102	
	C	-0.100	-0.10
	D	-0.098	
	E	-0.096	
2	A	-0.094	
	B	-0.092	
	C	-0.090	-0.09
	D	-0.088	
	E	-0.086	
3	A	-0.084	
	B	-0.082	
	C	-0.080	-0.08
	D	-0.078	
	E	-0.076	
.			
.			
.			
11	A	-0.004	
	B	-0.002	
	C	0.000	0.00
	D	+0.002	
	E	+0.004	
.			
.			
.			
21	A	+0.096	
	B	+0.098	
	C	+0.100	+0.10
	D	+0.102	
	E	+0.104	

In this study the Class Size is varied from 5 students per class, i.e., 25 students per school to 23 students per class, i.e., 115 students per school, in increments of two students per class. These two variables, Effect Size and Class Size, give us two important factors that influence the outcome of HLM.

The HLM software used is HLM2 Version 3.01 from SSI Scientific Software International with Visual Basic as the front-end

Results

For each combination of Effect Size and Class Size, 100 simulations were carried out and from the Empirical Bayes residuals of each simulation, the ranking of the schools were obtained. Thus we have 100 rankings of the schools, from each simulation. For each ranking, the correlation was found with the true ranking, and a correlation above a defined cutoff was assumed to give a satisfactory ranking of the schools for that simulation.

To observe the outcomes from the simulations, Table 2 below displays the results from six simulations for the Effect Size of 0.004 and a Class Size of 15. The columns marked HIGH are correlated at 0.97 or higher, columns marked MED are correlated at 0.91 and columns marked LOW are correlated at 0.86 or lower with the true rankings.

Table 2. Ranking results for selected simulations for Effect Size of 0.004 and Class Size of 15.

SLN	HIGH	HIGH	MED	MED	LOW	LOW	AVE
1	001	004	003	003	003	001	001
2	002	001	002	004	004	005	002
3	003	002	001	005	001	003	003
4	004	003	005	002	005	008	004
5	005	006	004	001	008	009	005
6	008	005	007	007	006	006	006
7	007	009	006	009	011	004	007
8	006	007	010	008	002	002	008
9	009	008	012	006	010	011	009
10	010	011	009	010	014	012	010
11	012	010	018	011	006	007	011
12	011	012	008	013	009	010	012
13	013	013	013	021	012	018	013
14	016	014	016	012	016	021	014
15	015	018	011	014	017	014	015
16	020	016	014	017	019	013	016
17	014	015	017	016	015	016	017
18	018	019	020	018	018	020	018
19	017	017	015	015	021	017	019
20	019	020	019	019	020	015	020
21	021	021	021	020	013	019	021
Correlation	0.97	0.97	0.91	0.91	0.86	0.85	1.00

For testing purposes, for each set of 100 simulations, the average Empirical Bayes residuals for the schools were calculated and the ranking resulting from these averages were recorded. The column marked AVE is the ranking of these average Empirical Bayes residuals for the 100 simulations. This ranking is identical with the true ranking, indicating that the simulation on the average yield true rankings and that the simulation technique yields valid results. In Table 2, schools that were not properly ranked by HLM

are in bold and schools whose ranking from HLM differs by more than two positions are shaded.

The most glaring observation was that not a single simulation run, irrespective of Class Size and Effect Size was able to rank the schools in the exact order. From the 10,000 simulations carried out for all the combinations of Effect Sizes and Class Sizes in this study, the best simulation result had two schools switching ranks in the middle of the distribution. If instability is defined as a ranking of a school over two positions from its true rank, then a characteristic that was consistently observed was the instability of the rankings at the middle of the distribution. And as the Effect Size and Class Size is decreased this region of instability tend to widen on either side of the median. This is an important factor that should be considered when classifying schools into above average and below average using hierarchical linear modeling. This phenomenon will occur at any boundary point when grouping takes place, but in classifying schools the significance of being below average and above average is of more importance than being in the first quartile or second quartile.

For each combination of Effect Size and Class size, after determining what a satisfactory ranking is, the number of successful rankings out of 100 simulations is determined. The Tables 3 through 11 below display the success rates of the simulations for different values of cut-off correlations. The combinations of Effect Size and Class Size that yield success rates below 95% are shaded. There are many possible reasons why different cut-off correlations would be selected. In areas other than school effectiveness analysis, where the interest is to observe similarities in effectiveness, a researcher can by choice select a lower or higher cut-off.

Table 3. Success rates of the simulations for a cutoff of $\rho = 0.90$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01
0.002	0.02	0.02	0.00	0.00	0.01	0.02	0.02	0.11	0.16	0.05
0.003	0.02	0.06	0.04	0.13	0.17	0.22	0.23	0.52	0.60	0.56
0.004	0.08	0.12	0.24	0.42	0.53	0.72	0.80	0.75	0.98	0.95
0.005	0.19	0.36	0.61	0.80	0.89	0.94	0.98	1.00	1.00	1.00
0.006	0.44	0.64	0.93	0.97	0.99	1.00	0.99	1.00	1.00	1.00
0.007	0.64	0.92	0.98	0.99	1.00	0.99	1.00	1.00	1.00	1.00
0.008	0.83	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.009	0.90	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4. Success rates of the simulations for a cutoff of $\rho = 0.91$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01

0.002	0.02	0.02	0.00	0.00	0.01	0.00	0.02	0.04	0.14	0.02
0.003	0.02	0.04	0.01	0.11	0.13	0.14	0.19	0.46	0.43	0.42
0.004	0.02	0.07	0.19	0.31	0.40	0.63	0.70	0.69	0.97	0.91
0.005	0.11	0.28	0.53	0.68	0.78	0.89	0.95	0.99	1.00	1.00
0.006	0.36	0.53	0.88	0.93	0.95	1.00	0.99	1.00	1.00	1.00
0.007	0.53	0.89	0.97	0.99	1.00	0.99	1.00	1.00	1.00	1.00
0.008	0.78	0.97	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.009	0.87	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 5. Success rates of the simulations for a cutoff of $\rho = 0.92$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01
0.002	0.02	0.02	0.00	0.00	0.01	0.00	0.02	0.03	0.12	0.00
0.003	0.01	0.03	0.00	0.07	0.06	0.11	0.11	0.37	0.29	0.34
0.004	0.00	0.06	0.11	0.15	0.34	0.48	0.59	0.62	0.92	0.83
0.005	0.04	0.20	0.38	0.58	0.71	0.80	0.88	0.96	0.99	0.99
0.006	0.21	0.37	0.71	0.86	0.92	0.99	0.99	1.00	1.00	1.00
0.007	0.48	0.79	0.92	0.99	1.00	0.99	1.00	1.00	1.00	1.00
0.008	0.68	0.90	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.009	0.82	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 6. Success rates of the simulations for a cutoff of $\rho = 0.93$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01
0.002	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.01	0.09	0.00
0.003	0.01	0.01	0.00	0.02	0.04	0.08	0.04	0.27	0.23	0.20
0.004	0.00	0.02	0.08	0.07	0.23	0.36	0.46	0.48	0.83	0.68
0.005	0.01	0.12	0.24	0.39	0.55	0.69	0.79	0.80	0.98	0.94
0.006	0.14	0.33	0.61	0.79	0.85	0.98	0.96	0.98	1.00	1.00
0.007	0.35	0.71	0.77	0.98	1.00	0.98	1.00	1.00	1.00	1.00
0.008	0.55	0.86	0.94	1.00	0.98	1.00	1.00	1.00	1.00	1.00
0.009	0.72	0.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 7. Success rates of the simulations for a cutoff of $\rho = 0.94$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01

0.002	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.09	0.00
0.003	0.01	0.00	0.00	0.02	0.02	0.04	0.02	0.13	0.15	0.17
0.004	0.00	0.02	0.06	0.02	0.16	0.23	0.31	0.41	0.71	0.53
0.005	0.01	0.08	0.16	0.26	0.41	0.51	0.62	0.69	0.95	0.84
0.006	0.06	0.19	0.43	0.69	0.71	0.94	0.93	0.95	1.00	1.00
0.007	0.26	0.52	0.66	0.96	0.93	0.96	1.00	1.00	1.00	1.00
0.008	0.44	0.79	0.87	0.98	0.98	1.00	1.00	1.00	1.00	1.00
0.009	0.57	0.88	0.96	0.98	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.92	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 8. Success rates of the simulations for a cutoff of $\rho = 0.95$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01
0.002	0.02	0.02	0.00	0.00	0.01	0.00	0.00	0.00	0.04	0.00
0.003	0.01	0.00	0.00	0.02	0.00	0.00	0.02	0.06	0.10	0.06
0.004	0.00	0.00	0.04	0.00	0.02	0.13	0.18	0.26	0.40	0.31
0.005	0.01	0.03	0.08	0.16	0.27	0.40	0.40	0.57	0.92	0.76
0.006	0.02	0.11	0.22	0.45	0.55	0.74	0.81	0.86	1.00	0.93
0.007	0.17	0.33	0.49	0.83	0.82	0.94	0.94	0.97	1.00	0.99
0.008	0.26	0.60	0.72	0.87	0.96	0.99	0.99	1.00	1.00	1.00
0.009	0.43	0.75	0.89	0.97	1.00	1.00	1.00	1.00	1.00	1.00
0.01	0.77	0.78	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00

Table 9. Success rates of the simulations for a cutoff of $\rho = 0.96$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01
0.002	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00
0.003	0.01	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.06	0.03
0.004	0.00	0.00	0.02	0.00	0.01	0.05	0.06	0.15	0.24	0.13
0.005	0.00	0.00	0.01	0.06	0.17	0.23	0.24	0.40	0.69	0.55
0.006	0.00	0.05	0.10	0.21	0.41	0.40	0.59	0.68	0.97	0.85
0.007	0.08	0.18	0.23	0.55	0.65	0.80	0.80	0.88	1.00	0.97
0.008	0.15	0.33	0.50	0.76	0.90	0.93	0.96	0.96	1.00	1.00
0.009	0.23	0.47	0.77	0.89	0.95	0.98	1.00	1.00	1.00	1.00
0.01	0.48	0.59	0.91	0.97	0.99	1.00	1.00	1.00	1.00	1.00

Table 10. Success rates of the simulations for a cutoff of $\rho = 0.97$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01

0.002	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.003	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.04	0.01
0.004	0.00	0.00	0.01	0.00	0.01	0.02	0.02	0.09	0.11	0.08
0.005	0.00	0.00	0.00	0.01	0.05	0.07	0.12	0.22	0.36	0.37
0.006	0.00	0.02	0.02	0.08	0.17	0.21	0.32	0.41	0.82	0.66
0.007	0.03	0.08	0.09	0.30	0.40	0.57	0.58	0.66	0.98	0.86
0.008	0.03	0.14	0.28	0.31	0.67	0.79	0.86	0.82	1.00	0.96
0.009	0.14	0.27	0.56	0.65	0.82	0.94	0.97	0.96	1.00	0.99
0.01	0.25	0.30	0.70	0.87	0.94	0.97	1.00	0.99	0.99	1.00

Table 11. Success rates of the simulations for a cutoff of $\rho = 0.98$

Effect Size	Class Size									
	5	7	9	11	13	15	17	19	21	23
0.001	0.04	0.05	0.03	0.03	0.04	0.00	0.01	0.01	0.00	0.01
0.002	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.003	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.004	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.02	0.01
0.005	0.00	0.00	0.00	0.00	0.00	0.02	0.03	0.03	0.10	0.11
0.006	0.00	0.00	0.00	0.01	0.02	0.06	0.13	0.11	0.33	0.30
0.007	0.01	0.01	0.03	0.08	0.12	0.26	0.20	0.29	0.64	0.51
0.008	0.02	0.04	0.09	0.10	0.31	0.50	0.47	0.52	0.95	0.79
0.009	0.00	0.07	0.25	0.38	0.49	0.56	0.78	0.65	0.99	0.92
0.01	0.03	0.12	0.33	0.51	0.60	0.79	0.86	0.86	0.90	0.97

For example, on Table 10, for a Class Size of 5 and Effect Size of 0.001, a success rate of 0.04 indicates that 4% of the simulations ranked the schools above the cut-off correlation of 0.97.

Conclusions

The aim of this simulation study was to examine the validity of the use of hierarchical linear modeling (HLM) in school effectiveness analysis. As such the conclusions that are needed are under what conditions can Hierarchical Linear Modeling be successfully used in school effectiveness analysis to rank schools. The most important of all condition that is needed is what is the minimal n , Class Size, effectively the number of students in a school, that is required to successfully calculate school rankings given that we know the Effect Size that distinguish any two schools.

From the simulation results and with the predetermined criteria of what a good ranking is, we can determine when HLM will yield accurate results. After deciding what level of correlation we want with the true ranking, from Tables 3 to 11, we can determine the Effect Sizes and Class Sizes that are necessary to yield confident results. For example, if it is decided that a cut-off correlation of 0.90 is satisfactory, if the Effect Size and Class

Size combination is anywhere in the unshaded region, HLM will produce satisfactory rankings with 95% confidence.

As expected, with higher cutoffs for the correlations the feasible region where HLM will yield satisfactory results diminishes. There is also a distinctive pattern observed for successful ranking from the above tables. The Class Size and Effect size are both equally important for successful school ranking for the range of Class Sizes and Effect Sizes discussed in this paper. For very high levels of cutoff correlations, i.e., expectations that HLM will produce almost exact rankings, the necessary Effect Size and Class Size combination goes beyond practicality. This is observed when the correlation cut-off is increased from 0.97 to 0.98, the feasible region where HLM ranking can be used with a 95% confidence has diminished significantly (see Tables 10 and 11). It can also be concluded that at cutoff of 0.98, HLM ranking will not be suitable at all.

The simulated student scores in this study were standardized normal. The smallest Effect Size considered is 0.001, which is 0.10% of the standard deviation of the generated student scores. If a mean school gain is calculated by averaging the students pre-test post-test gains, the median school has an average gain of zero and the next best school has an average gain of 0.001. What is the significance of Effect Size in this context? For Class Size of 5, the mean school gains are distributed normal with mean zero and standard deviation of 0.20. The resulting Effect Size used to bias the students' post-test scores is now 0.50% of the standard deviation. This signifies that when considered at the school level, the mean of the distributions of any two consecutively ranked schools are not as close as they appear. In practice, analyzing the average gains of the schools independent of any other student level and school level co-variates may be a technique which could be employed to estimate the Effect Size present in the data and thus the minimum Class Size required to rank schools.

In the course of this simulation study, only 100 simulations per combination of Effect Size and Class Size were carried out. Is this sufficient? Ideally, it would have been suitable to carry out a higher number of simulations for each Effect size and Class size combination. It takes approximately sixty minutes to do 100 simulations. Multiplying this by the ten Effect Sizes and the ten Class Sizes, this simulation study took approximately one-hundred hours of CPU time. If the number of simulations were increased to 1000, i.e., ten folds, the amount of CPU time necessary would be approximately one-thousand hours. The authors propose to narrow the Effect Size and Class Size combinations and carry out more simulations per combination in the future. Future consideration also include a power analyses when unbalanced groups are considered and studies of results when second level characteristics, i.e., school level covariates, are introduced into the model.

In conclusion, can HLM rank schools? The answer certainly is yes, provided that the data fulfills the assumptions of normality and the Class Size, i.e., the number of students per school, is relatively large with respect to the effects that distinguish any two schools.

References:

- Anderson, T. W., (1971). *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.
- Aiken, L. S. and West, S. G. (1991). *Multiple Regression: Testing and Interpreting Interactions*. Newburg Park: Sage.
- Bock, R. Darrell, (1989). *Multilevel Analysis of Educational Data*. San Diego, CA: Academic Press.
- Bryk, A. S., and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newburg Press, California; Sage Publications.
- Bryk, A. S., Raudenbush, S. W., Seltzer, M. and Congdon, R. (1988). *An Introduction to HLM: Computer Program User's Guide (2nd ed.)* Chicago, Ill: University of Chicago.
- Cohen, J. and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, New Jersey: Laurence Erlbaum.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm*. Journal of the Royal Statistical Society, Series B, 39, 1-8.
- Darlington, R. B. (1990). *Regression and Linear Models*. New York: McGraw-Hill.
- Goldstein, H., (1987). *Multilevel Models in Educational and Social Research*. New York, NY: Oxford University Press.
- Laird, N. M. and Ware, H. (1982). *Random-Effects Models for Longitudinal Data*. Biometrics, 38, 963-974.
- De Leeuw, J. and Kreft, I. G. G. (1995). *Questioning Multilevel Models*. Journal of Educational and Behavioral Statistics, Summer 1995, Vol. 20, No. 2, pp. 171-189.
- Longford, N. T. (1987). *A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects*. Biometrika, 74 (4), 817- 827.
- Mendro, R. L., Webster, W. J., Bemby, K., and Orsak, T. H. (1994). *An Application of Hierarchical Linear Modeling In Determining School Effectiveness*. Rocky Mountain Educational Research Association, Phoenix, Arizona.
- Millman, J. (ed.) (1981). *Handbook of Teacher Evaluation*. Beverly Hills, California, Sage Publications.
- Rosenberg, B. (1973). *Linear Regression With Randomly Dispersed Parameters*. Biometrika, 60, 61-75.
- Raudenbusch, S. W. (1995). *Reexamining, Reaffirming, and Improving Application of Hierarchical Models*. Journal of Educational and Behavioral Statistics, Summer 1995, Vol. 20, No. 2, pp. 210-220.
- Saka, T. (1984). *Indicators of School Effectiveness: Which are the Most Valid and What Impacts Upon Them?* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA: (ERIC No. ED 306277).
- Snijders, T. B. and Bosker, R. J. (1993). *Standard Errors and Sample Sizes for Two-Level Research*. Journal of Educational Statistics, Fall 1993, Vol. 18, No. 3, pp. 237-259.

Webster, W. J., Mendro, R. L., Bembry, K. and Orsak, T. H. (1995). *Alternative Methodologies for Identifying Effective Schools*. Distinguished Paper Session. American Educational Research Association, San Francisco, CA. ERIC EA 027 189.

Webster, W. J., Mendro, R. L., Orsak, T. H., and Weerasinghe, D. (1996). *The Applicability of Selected Regression and Hierarchical Linear Models To The Estimation of School and Teacher Effects*. Paper Session. American Educational Research Association, New York, NY.

Weerasinghe, D., Orsak, T. H., and Mendro, R. L., (1997). *Value Added Productivity Indicators: A Statistical Comparison of The Pre-Test/Post-Test Model and Gain Model*. Southwest Educational Research Association, Austin, Texas, January 1997.