Comparing OLS and HLM Models and the Questions They Answer: Potential Concerns for Type VI Errors

David Newman Cleveland State University
Isadore Newman Florida International University
James Salzman Ohio University

Hierarchical Linear Modeling (HLM) has become an important analytical tool in a number of fields of study and many of the educational journal articles published during the last decade have used this technique. In this study, the authors proposed that a clearer understanding of what HLM was truly testing could be facilitated if the researcher constructed multiple linear regression (MLR) models to reflect the research questions that the corresponding HLM was intending to test. The authors compared the results of HLM and MLR models on a sample data set, discussed the advantages and disadvantages of each technique, and provided some examples of the increased risk of committing Type VI errors in running an HLM model without understanding the underlying question that was being asked. Further issues that pertained to researchers' obligations to differentiate between statistical analysis and research design were discussed in the areas of centering, interaction, and fixed, mixed, and randomized effects.

Hierarchical Linear Models (HLM), which are frequently referred to as multilevel models, are most appropriately and effectively used when variables tend to be nested within other variables. For example, individual students may be nested within classes and classes could be nested within schools. A number of researchers (Field, 2009; Kreft, 1996; Morris, 1995; Mundfrom & Schultz, 2002; Raudenbush & Bryk, 2002; Tabachnick & Fidell, 2007) have indicated that HLM is superior to ordinary least squares (OLS/General Linear Models) because HLM theoretically produces appropriate error terms that control for potential dependency due to nesting effects, while OLS does not. An additional argument favoring the use of HLM is that it is a generalization of OLS, which better handles continuous variables that reflect randomized effect designs, and, therefore, HLM produces more accurate error terms and Type I error rates (Mundfrom & Schultz; Raudenbush, 2009; Raudenbush & Bryk). A good part of the cited advantages for HLM is related to the situations in which the intraclass correlations, which is the between group effect divided by the total effect, departs from zero with $\rho = \tau_{00}/(\tau_{00} + \sigma^2)$. If the correlation is zero, there seems to be less advantage to using HLM because there is no interclass correlation.

Method

It is the contention of this paper that the stated advantage for using HLM instead of the traditional OLS/Multiple Linear Regression (MLR) models, is that the MLR models typically were not written to most appropriately reflect the research questions in nested designs. We believe this is frequently due to the failure to incorporate person vectors (McNeil, Newman & Kelly, 1996; Pedazur, 1982; Williams, 1987). A person vector is binary coded so that if a dependent score comes from a particular unit, such as a person who is repeated, a school, or a district, etc., it is coded as 1 and coded as a 0 if it does not come from that unit. The person vector coding, thus, controls for the variability due to the unit. Without this type of coding, there tends to be a confounding or nesting effect that leads to an inconsistency between the research question and designing an appropriate model to test the question. We refer to this inconsistency as a Type VI Error (Newman, Fraas, Newman & Brown, 2002; Tracz, Nelson, Newman & Beltran, 2005; Tracz, Newman, Nelson, & Dellran, 2004).

Results and Discussion

OLS Models

The following examples were run to demonstrate how different models could be written in OLS/MLR to reflect different research questions. For example, if one were interested in determining if a treatment predicts Ohio Achievement Test (OAT) scores, the following three models could be written to test this question:

Table 1. Model 1: $OAT = a_0u + a_1 (Treatment) + Error$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.038[a]	.001	.001	28.791	.001	2.679	1	1819	.102

a. Predictors: (Constant), TX

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	404.790	.954		424.360	.000
	TX	2.209	1.349	.038	1.637	.102

a. Dependent Variable: OAT_09

This model would be an over simplification of the modeling of the question of interest if the treatment were nested within schools. The results of this model in Table 1 indicate that $a_1 = 2.209$ with a $p = .102$; meaning the treatment was not statistically significant at alpha = .05. A more appropriate model to reflect the question of interest when treatment is nested would be:

Model 2: $OAT = b_{01} + b_{37}(School\_1) + b_{38}(School\_2) + b_{39}(School\_3) + \quad b_{69}(School\_34) + E$
Model 3: $OAT = b_0u + b_1 (Treatment) + b_2 (School\ 1) + \ldots + b_N (School\ N) + E$

Table 2. Model 2 & 3 Summary.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
2	.385[a]	.148	.132	26.841	.148	9.117	34	1786	.000
3	.388[b]	.151	.134	26.801	.003	6.376	1	1785	.012

a. Predictors: (Constant), School_34, School_28, School_14, School_23, School_33, School_15, School_10, School_6, School_8, School_30, School_11, School_16, School_32, School_31, School_22, School_24, School_19, School_4, School_5, School_25, School_7, School_12, School_21, School_29, School_2, School_26, School_20, School_13, School_17, School_18, School_9, School_27, School_1, School_3

b. Predictors: (Constant), School_34, School_28, School_14, School_23, School_33, School_15, School_10, School_6, School_8, School_30, School_11, School_16, School_32, School_31, School_22, School_24, School_19, School_4, School_5, School_25, School_7, School_12, School_21, School_29, School_2, School_26, School_20, School_13, School_17, School_18, School_9, School_27, School_1, School_3, TX

The first model is simply testing to see if there is a relationship between people who had the treatment and the OAT scores. Testing the second model against Model 3 is considerably different in that it is testing to see if the treatment has an effect independent of school differences. Please note that b_1 is now 4.016 compared to a_1 in Model 1, which is 2.209. As seen in Table 2, when Model 2 is tested against Model 3, there is a statistically significant difference ($p = .012$). By adding in the school vectors, the model more accurately reflects the question of interest. As one can see, the number of independent vectors has increased by the number of schools minus one (K-1) for both the full and restricted models. Table 3 shows that when the random nature and violation of independence of errors are accounted for, the fixed effect for treatment now becomes statistically significant with $p = .012$ and a regression coefficient = 4.016.

Table 3. Models 2 & 3 Coefficients Table

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
2	(Constant)	415.638	3.524		117.932	0
	School_1	0.798	4.224	0.007	0.189	0.85
	School_2	-7.527	4.884	-0.048	-1.541	0.123
	School_3	6.033	4.121	0.059	1.464	0.143
	.					
	.					
	School_33	5.232	6.614	0.02	0.791	0.429
	School_34	-10.683	4.831	-0.069	-2.211	0.027
3	(Constant)	415.569	3.519		118.086	0
	School_1	-2.514	4.416	-0.023	-0.569	0.569
	School_2	-10.645	5.031	-0.068	-2.116	0.034
	School_3	5.314	4.124	0.052	1.288	0.198
	.					
	.					
	.					
	School_33	2.333	6.703	0.009	0.348	0.728
	School_34	-14.143	5.014	-0.092	-2.821	0.005
	TX	4.016	1.59	0.07	2.525	0.012

a. Dependent Variable: OAT_09

It is our contention that HLM has been frequently compared to MLR models that are often incorrectly written because they do not contain what we call "person vectors" or in the above case "school vectors." If one looks at the HLM Level 1 and Level 2 models presented below, it is apparent that they are much more similar to Model 2 above (see the output below that reflects this).

Multileveled Modeling

The Multileveled Models/Linear Mixed Models (LMM) reflect the same question that was tested by the regression models containing school vectors. In this case, the actual mixed model is created by substituting B_{0j} with B_0+u_{0j} and B_{1j} is replaced with y_{10}, but the error u_{1j} is not included since the differential treatment effects across schools are not being tested. This fourth set of models represents the LMM and the combined mixed model.

Model 4: LMM

Level 1 (Student): $OAT\_9 = B_{0j}+B_{1j}(TX_i)+r_{ij}$

Level 2(School) $B_{0j} = y_{00} + u_{0j}$

$B_{1j} = y_{10} +$

Mixed Model: $OAT\_9 = y_{00}+u_{0j}+y_{10}(TX)+r_{ij}$

Table 4 displays the estimates of the overall goodness of fit for this model looking at the -2 Restricted Log Likelihood statistic. Assessing fit and comparing different multilevel models becomes necessary in calculating the chi-square statistic.

χ^2= (-2 Restricted Log Likelihood$_{old}$) (-2 Restricted Log Likelihood$_{New}$)

df= Number of Parameters$_{old}$ – Number of Parameters$_{New}$

Table 4. Goodness of Fit Estimates

Information Criteria[a]

-2 Restricted Log Likelihood	16944.349
Akaike's Information Criterion (AIC)	16950.349
Hurvich and Tsai's Criterion (AICC)	16950.362
Bozdogan's Criterion (CAIC)	16969.810
Schwarz's Bayesian Criterion (BIC)	16966.810

The information criteria are displayed in smaller-is-better forms.

a. Dependent Variable: OAT_09.

The other goodness of fit estimates (e.g., AIC, AICC, CAIC, BIC) all adjust for different model complexities such as the number of parameters in the model and the sample size.

The Type III Test of Fixed Effects gives the overall fixed effect for TX and schools. In this case, only TX is of interest because schools are considered to be random. It should be noted at this point that Treatment is a fixed effect. It is also possible to have Treatment as a random effect depending on how it is operationally defined. In this case, the student either received treatment or did not. As can be seen, the p-value for the fixed effect of TX ($p = 0.012$) is the same as the p-value calculated when testing Model 2 against Model 3.

Table 5. Fixed Effects Coefficients

Type III Tests of Fixed Effects[a]

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	1785	313550.052	.000
TX	1	1785	6.376	.012
s_irn_2009	34	1785	9.241	.000

a. Dependent Variable: OAT_09.

Table 6 displays the coefficients of the fixed effects. We were not interested in the fixed effects of schools because not all of the schools were included in this illustration. As one can see from Table 6, the partial regression coefficient for TX is = -4.0159, which is again the same as that in the regression model with school vectors.

As noted earlier, the question of interest is "Does treatment account for a significant proportion of unique variance in predicting OAT independent of school differences?" Students who received the treatment scored, on average, 4.0159 points higher on the OAT than their comparison group. Even though this is a small difference ($R^2_{change}=0.003$), when either the LMM or regression is specified correctly, the power to detect these types of differences in nested design data becomes possible and equivalent. *Type VI Error*

A Type VI Error is a catchall concept that describes the inconstancy between the research question of interest and the statistical model (Newman et al., 2002; Tracz et al., 2005; Tracz et al., 2004). Too often these inconsistencies are overlooked leading researchers to make incorrect inferences. In addition to specifying the appropriate model, here are some other areas related to multilevel modeling that may be related to Type VI Errors.

Table 6. Fixed Effects Estimates

Estimates of Fixed Effectsb

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	419.584607	3.850561	1785	108.967	.000	412.032526	427.136689
[TX=.00]	-4.015916	1.590371	1785	-2.525	.012	-7.135100	-.896732
[TX=1.00]	0	0	.	.	.	.	.
[s_irn_2009=257.00]	-2.514425	4.416500	1785	-.569	.569	-11.176479	6.147629
[s_irn_2009=352.00]	-10.644815	5.030888	1785	-2.116	.034	-20.511864	-.777766
[s_irn_2009=388.00]	5.314262	4.124444	1785	1.288	.198	-2.774985	13.403509
[s_irn_2009=22012.00]	-11.446891	5.787912	1785	-1.978	.048	-22.798689	-.095094
[s_irn_2009=23192.00]	2.332588	6.703082	1785	.348	.728	-10.814125	15.479301
[s_irn_2009=23283.00]	-14.143284	5.014410	1785	-2.821	.005	-23.978016	-4.308553

Centering

Centering is used most frequently when running mixed models or HLM. It is simply subtracting the mean from each score so that the mean of the distribution becomes zero. The choice to center is not a simple mathematical or statistical decision. It should be based upon the researcher's question of interest and/or theoretical position. There are three major decisions one has to make about centering: 1). Should one center? 2). If centering, should grand mean centering be used? 3). Should one use group mean centering? Grand mean centering is generally preferred over group mean centering (Burton, 1993; Hoffman & Gavin, 1998; Kreft, de Leeuw, & Aiken, 1995). Sarkisian (2007) takes the position that the original metric should never be used if the value of zero is not meaningful, and finds that there is a lack of precision in estimating the intercept in HLM when one does not center. According to Field (2009), centering is not an easy decision. It requires an understanding of the data and the analysis. Field also suggests centering may be a useful way to ameliorate the problem of multicollinearity between independent variables, especially when the independent variable does not have an interpretable zero value. Field points out that it is important to note that when using the group mean centering approach, the group mean should be considered a second level variable whenever group effects are not of interest. This situation is encountered frequently when an independent variable, such as time, is of interest.

If the researcher is interested in the relative position of the subject with regard to the treatment group mean, then group mean centering should be used. If, on the other hand, the researcher is interested in the absolute value of the independent variable (predictor variable), then grand mean centering should be used. When one does grand mean centering, the intercept becomes the adjusted grand mean. This adjustment obviously does not have any effect on the slopes. When group mean centering is used, the intercept is interpreted as the mean of each group. Group mean centering may change the meaning of the coefficient so that it becomes difficult to interpret because the mean values are subtracted from different sets of raw data. Some researchers will even center group mean binary variables; one needs to keep in mind that with group centered predictor variables, only person level effects are estimated. Choosing to center or not to center, and determining whether to use grand or group means, relates to Type VI Error because each decision will affect the statistical model that will differentially reflect the research question of interest.

Interaction

The classical definition of interaction is the differential effect across an area of interest (non-equal slopes) over and above the main effect (or controlling for the main effect). If looking at HLM to see if the second level accounts for a significant proportion of variance, one is looking at a differential effect across the area of interest, but it is not over and above the main effects. In other words, only the multiplicative slope differences are being tested, but not the slope differences independent of the main effects. Comparing the first level to second level HLM models is very similar to traditional interaction, but because it does not include the main effects, the results could be different.

Adequacy of Sample Size

Another important issue that is too often overlooked when using HLM is the adequacy of the sample size, especially as it relates to the higher-level variables in a model. Kreft (1996) found that to have sufficient power one needs at least 30 groups with 30 subjects per group, or 60 groups with 25 replicates per group, or 150 groups having 5 replicates per group. Kreft's simulation data suggests that the number of groups is more important than the number of observations for statistical power. This potential lack of power, from not having enough groups, has implications for detecting interaction between levels. Hox (1995) and Hox and Maas (2001) have similar findings related to adequacy of sample size. They found that $N < 20$ is insufficient at the higher levels, and if these higher-level variables are crucial to the structural model, then N should be > 100. These results are not consistent with the position taken by Raudenbush and Bryk (2002) who believe a Bayesian estimation approach allows for smaller N.

Research Design (Fixed Effects and Randomized Effects)

One of the major issues related to HLM and OLS/MLR is that the researcher has to determine if the design consists of fixed, mixed, or random effects. It is important to know the nature of these effects so that one can select the appropriate error terms. If the correct error term is not selected, the researcher cannot determine the correct error rates for the tests of significance.

Fixed effects occur when the variables of interest are assumed to not be randomly selected and no generalizations are going to be made beyond the variables being tested. For example, if there are three treatments that a researcher is interested in testing to see if they have a differential effect, then only those specific three treatments would be tested. This variable is fixed. In another example, imagine that a researcher is interested in the effects of a range of drug doses and randomly selects three doses to be representative of the entire range. Because the researcher wants to infer to the whole range of doses from which the samples are drawn, the three levels of the selected doses are considered to be random effects. In mixed effects, one must have at least two independent variables with at least one fixed and one random.

When building regression or hierarchical linear models, different error terms are used to test for statistical significance depending upon whether the variables are fixed or random. Being able to determine if the variables are fixed or random is important in testing models because they have different assumptions. In our opinion, one important issue here is conceptual; not statistical. As an example, assume a researcher is interested in having drug dosage as a continuous variable, and interested in generalizing to the whole range of dosages (i.e., think of this as a variable on the X axis of a graph). In this scenario, it would appear that dosage is a continuous variable. If, however, the researcher were to take drug dosage and was only interested in generalizing to three categories (i.e., small, medium, or large doses) based upon some predetermined decision rule, then the variable dosage would have changed from a continuous (random) variable to a categorical (fixed) variable. Obviously, with the fixed variable the researcher is specifically addressing whether there is a difference between the small, medium, and large dosage levels and not attempting to generalize to the range of doses. The robustness of violations to the underlying assumptions of the fixed and randomized models, as they relate to the accuracy of the tests of significance, is considerable. The fixed model, especially when the design is balanced, is more robust than the randomized model.

Conclusion

The authors contend that there is nothing more important than understanding the relationship between the question of interest, the data, and the analysis. There is no computer program that is capable of doing the researcher's thinking for him or her. Therefore, it could be very misleading to use default options on computer programs or to use very sophisticated computer programs that have algorithms that are virtually black boxes and are too often not understood by the researcher. HLM is being widely used to analyze data, but we are concerned about how well these models are understood and appropriately interpreted. This concern is especially true when considering that stability is heavily based upon the number of replicates at the higher-order levels.

HLM modeling for the above example frequently takes into consideration the differential effect of treatment for different schools (a "type of interaction"). However, this question was not the one that was posed in the above OLS models. One could run models to reflect interaction, but that would go beyond the purpose of this paper. We have demonstrated that when models were written to reflect the question of interest ("Was there an overall different effect due to treatment, independent of schools?"), the results of OLS and HLM were virtually identical.

To complicate matters, there is an option in HLM to run treatment as a fixed or random effect. If treatment is thought of as a random effect, (this choice would not make sense in the above example), and if HLM were run using treatment as a random effect, the results would be different from those obtained using OLS modeling. Additionally, various "experts" in HLM suggest different approaches for writing the HLM models in their reference books. These diverse approaches may produce dissimilar results (see Bickel, 2007; Field, 2009; Raudenbush & Bryk, 2002). The choice to use random or fixed effects, the type of centering selected, etc. needs to be contingent upon the researcher's understanding of the data and the purpose of the research. While there are advantages to using HLM, researchers must be aware that it is not always the most efficacious procedure. One must select the method that will most accurately reflect the research question in the simplest way. This selection is analogous to the concept of parsimony.

References

Bickel, R. (2007). *Multilevel analysis for applied research: It's on regression.* New York: The Guilford Press.

Burton, B. (1993). *Some observations on the effect of centering on the results obtained from hierarchical linear modeling.* Washington, D.C.: National Center for Education Statistics, U. S. Department of Education.

Field, A. (2009). *Discovering statistics using SPSS.* Thousand Oaks, CA: Sage Publication.

Hoffman, D. A., & Gavin, M. B. (1998). Centering decisions in hierarchical linear models: Implications for research organizations. *Journal of Management, 24*, 623-641.

Hox, J. J. (1995). *Applied multi-level analysis: A basic, non-technical introductory text* (2nd ed.). Amsterdam, Netherlands: TT-Publikaties.

Hox, J. J., & Maas, C. J. M. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation Modeling, 8*, 157-174.

Kreft, I. G. G. (1996). *Are multi-level techniques necessary? An overview, including simulation studies.* Retrieved November 22, 1999, from http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research 30*, 1-21.

McNeil, K., Newman, I., & Kelly, F. J. (1996). *Testing research hypotheses with the general linear model.* Carbondale, IL: Southern Illinois University Press.

Morris, C. (1995). Hierarchical models for educational data- an overview. *Journal of Educational and Behavior Statistics, 20*, 190-200.

Mundfrom, D. J., & Schults, M. R. (2002). A monte carlo simulation comparing parameter estimates from multiple linear regression and hierarchical linear modeling. *Multiple Regression Viewpoints, 28*, 18-21.

Newman, I., Fraas, J., Newman, C., & Brown, R. (Fall, 2002). Research practices that produce Type VI Errors. *Journal of Research in Education, 12*, 138-145.

Pedazur, E. J. (1982). *Multiple regression and behavioral research: Explanation and prediction* (2nd ed.). New York: Holt, Rinehart & Winston.

Raudenbush, S. W. (2009) Analyzing effect sizes: Random effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis.* (pp. 295- 315). New York: Russell Sage Foundation.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Sarkisian, N. (2007). *HLM model building strategies: Class notes.* Retrieved April 2, 2010, from www.sarkisian.net/sc705/september6.pdf

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

Tracz, S., Nelson, L., Newman, I., & Beltran, A. (2005). The misuse of ANCOVA: The academic and political implications of Type VI errors in studies of achievement and socioeconomic status. *Multiple Linear Regression Viewpoints, 31*, 16-21.

Tracz, S., Newman, I., Nelson, L., & Dellran, A. (2004, October). *How ANCOVA can be misused in studies of achievement and socioeconomic status: The academic and political implications of Type VI* *errors*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Williams, J. D. (1987). The use of nonsense coding with ANOVA situations. *Multiple Linear Regression Viewpoints 15*, 29-39.

Send correspondence to: David Newman
Cleveland State University
Email: d.o.newman@csuohio.edu