Multilevel Modeling: Clarifying Issues of Concern

David Newman
Florida Atlantic University

Isadore Newman
Florida International University

When using Hierarchical Linear Modeling (HLM) to analyze complex nested data, it is important to consider issues that affect the interpretation of the HLM outcomes. Alternative methods of accounting for the variance within the nested structures need to be considered if they better fit the research question of interest. One alternative method to HLM is Multiple Linear Regression (MLR)-Ordinary Least Squares solutions with person vectors. This study compares a number of sets of data that reflect interaction questions as well as nested designs. More specifically, eight issues that need to be considered when using HLM are discussed. These issues are: 1) advantages of HLM; 2 & 3) person vectors as it relates to nesting; 4) centering; 5) picking the appropriate error terms for fixed, random, and mixed effects; 6) understanding interaction and how it is tested; 7) sample size related to first and second level models; and 8) comparing similarities and differences between HLM and MLR with person vectors.

Hierarchical Linear Modeling (HLM) is the general case of Multiple Linear Regression (MLR) and a preferred statistical method for analyzing complex, nested data structures. While very powerful, if one does not understand where its use is advantageous and how and why to make necessary decisions in building the HLM models, it is possible to write statistical models that do not reflect the research question(s) of interest; thus, committing what is known as a Type VI error (Newman, Fraas, Newman, & Brown, 2002). The purpose of this study is to facilitate a better understanding of multilevel modeling by discussing eight topics that require consideration when using HLM. These topics are: 1) advantages of HLM; 2 & 3) person vectors as it relates to nesting; 4) centering; 5) picking the appropriate error terms for fixed random and mixed effects; 6) understanding interaction and how it is tested; 7) sample size related to first and second level models; and 8) comparing similarities and differences between HLM and MLR with person vectors.

Advantages of HLM within Subjects Nested Effects

The hierarchical linear family of models, 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, patients may be nested within hospital floors and floors may be nested within hospitals. A number of researchers (Field, 2009; Kreft, 1996; Morris, 1995; Mundform & Schultz, 2002; Raudenbush & Bryk, 2002; Tabachnick & Fidell, 2007) have indicated that HLM is superior to Ordinary Least Squares (OLS) - General Linear Model 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 that better deals with continuous variables, which reflect randomized effect designs, and; therefore, HLM produces more accurate error terms and Type I error rates (Mundform & Schultz; Raudenbush, 2009; Raudenbush & Bryk). An aspect of the cited advantages for HLM is related to the situations in which the interclass correlations (ICC), which is the between group effect divided by the total effect, depart from zero. There has been some argument that the closer the correlation gets to 0, the less advantage there is to using HLM because there is a low level of organizational correlation. Lee (2000) suggested that if the ICC is less than 10% (i.e., $r^2 \leq .1$ or $r \leq .316$) of the total variance, one can assume that there is no meaningful nesting effect. This rule of thumb is still a topic of debate.

In addition to the production of appropriate error terms, HLM is the one of the most flexible techniques in reflecting change over time for individual subjects (i.e., HLM also known as linear mixed modeling and growth modeling). Depending on the nature of the individual change function, parameter estimates are calculated for the intercept, slope and, if necessary, curvature to create the best predictive model fit for the data. Because HLM shows longitudinal changes as a function of time, the time variable can be centered at the baseline. Thus, in the examination of model growth parameters, the intercept will represent the level of outcome of the baseline assessments and the slope will indicate the rate at which the level is changing. Additionally, unlike the traditional repeated measures analysis of variance, HLM does not delete cases from the analysis because of missing time points. Also, HLM allows for flexibility in the distance between individual measurements (McCulloch & Searle, 2001; Singer & Willett, 2003).

Person Vectors and Nesting. An alternative way to account for the variance of a nested structure in MLR is to use person vectors (McNeil, Newman, & Fraas, 2012; Williams, 1987). A person vector is simply a vector that consists of ones and zeros that identifies which scores on the criterion variable are associated with each person or organized structure such as a specific floor in a hospital. If we had a criterion variable that had three scores from Person 1 and three scores from Person 2, it would look like the following example in Table 1.

Table 1. Person Vector Example

ID	Y	P1	P2
1	100	**1**	0
1	105	**1**	0
1	107	**1**	0
2	85	0	**1**
2	92	0	**1**
2	98	0	**1**

As can be seen from the above matrix, the first three scores came from Person 1 and the second three scores came from Person 2. These vectors allow one to account for variance due to individual differences.

An alternative technique using MLR and person vectors that is being used in educational and medical settings is repeated measures regression discontinuity analysis (RD). RD can test multiple time points against each other. For instance, if there were four time points that were being used to measure a person's pain, it would be possible to test both slope and intercept differences between time 1 versus time 2, time 3 versus time 4, time 1 versus time 3, and time 1 versus time 2. Using this technique, one can test the pre-test and the post-test slope differences; independent of individual differences. One can also test for function differences with the pre-test and the post-test. The following figure, Figure 1, is an example of slope differences between pre-test and post-test. It is possible to test individual time point differences as well as testing slope difference. Another name for this type of analysis is an interrupted time series (Campbell & Stanley, 1963).

The following models allow us to test for slope and intercept differences, while controlling for individual differences, by testing Model 1 against Model 2, where:

$\mathbf{S_1}$ = the slope for line 1 in the above figure

$\mathbf{S_2}$ = the slope for line 2 in the above figure

$\mathbf{U_1}$ = 1 if the subject's score on the criterion variable came from the pre-treatment or zero otherwise

$\mathbf{U_2}$ = 1 if the subject's score on the criterion variable came from the post-treatment or zero otherwise

Model 1: Restricted (R^2_1): $Y = a_cU_c + a_cS_c + Zb_1(P_1) + \ldots Zb_n(P_n) + E_1$

Model 2: Full (R^2_2): $Y = a_{01}U_1 + a_1S_1 + a_{02}U_2 + a_2S_2 + Zb_1(P_1) + \ldots Zb_n(P_n) + E_2$

We can also test for intercept differences, while controlling for individual and slope differences, by testing Model 2 against Model 3 based upon the following restriction: $a_{01} = a_{02}$

Model 3: $Y = a_{03}U_1 + a_4S_1 + a_5S_2 + Zb_1(P_1) + \ldots Zb_n(P_n) + E_3$

Looking at Figure 1, in a similar manner, we could test the restriction on the full model based upon: $(t_1 - t_3) = (t_4 - t_6)$. Another hypothesis could be based upon testing $t_3 = t_4$, independent of individual differences, etc. Furthermore, there are several other issues that need to be understood and decided upon when using either HLM or MLR with person vectors. These issues directly impact the research question of interest and if they are not handled appropriately, one is likely to make a Type VI error or when the statistical model does not align with the research question (Newman et al., 2002; Tracz, Nelson, Newman, Beltran, 2005; Tracz, Newman, Nelson, Dellran, 2004).

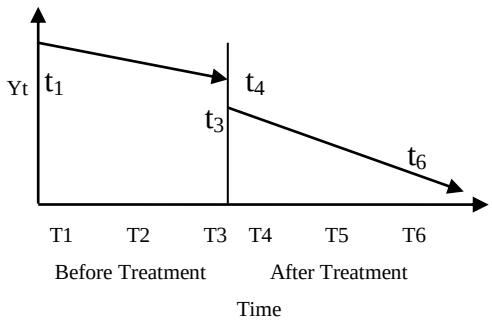

Figure 1. Testing Intercept and Slope Differences.

Centering. The first of the additional issues to be discussed is centering. Centering 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?; and 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. Sarkisian further 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 specific types of independent variables, especially when the independent variable does not have an interpretable zero value. Field also 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 frequently encountered 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 (i.e., 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; however, 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.

Error Terms, Research Design (Fixed Effects and Randomized Effects). *One of the major issues* related to HLM and OLS/MLR is that one has to determine if the design consists of fixed, mixed, or random effects. It is important to know the nature of these effects so the appropriate error terms can be selected. If the appropriate error term is not selected, the researcher cannot determine the correct error rates for the tests of statistical significance. ***Fixed effects*** occur when it is assumed that the variables of interest are not randomly selected and no generalizations are going to be made beyond the variables being tested. For example, if there are three treatments that one is interested in testing to see if they have a differential effect, then only those specific three treatments would be tested. Thus, this variable is fixed. In a different 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. Since 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 a ***mixed effects*** design, one must have at least two independent variables with one fixed and the other random (Kirk, 1968).

When building regression models, or hierarchical linear models, different error terms are used to test for significance depending upon whether the variables are fixed or random. This is important in model testing because fixed and random variables have different assumptions associated with them. In our opinion, one important issue here is conceptual and not statistical. As an example, assume one is interested in having drug dosage as a continuous variable and is 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, dosage is a continuous variable. However, if in looking at drug dosage, the researcher was only interested in generalizing to small, medium, or large doses based upon some predetermined decision rule, and he or she was only interested in those three categories, then the variable dosage 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 as operationally defined, and is 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.

Interaction. The test for interaction is the next important issue. The classical definition of interaction is the differential effect across an area of interest (i.e., non-equal slopes) over and above the main effect

(i.e., controlling for the main effect) (McNeil et al., 2012). In HLM, interaction is frequently inferred by comparing the first level to the second level model. 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 the second level HLM models is very similar to traditional, but since it does not include the main effects, the results could be different. It is important that researchers are aware of their research questions and how well-chosen models reflect their question of interest.

With regression, as suggested by McNeil et al. (2012), one can write three different types of interaction. These are: 1) interaction between categorical and categorical variables such as interaction between sex and treatment; 2) categorical and continuous interaction such as interaction between sex and IQ; and 3) continuous and continuous interaction such as interaction between IQ and Motivation. Each one of these will test the multiplicative effect over and above the additive effect.

Categorical –Categorical Interaction (Treatment and Sex)

Model 4 Full: $Y = a_0u + a_1T_1m + a_2T_1f + a_3T_2m + a_4T_2f + E_4$

Model 5 Restricted: $Y = a_0u + a_5T_1 + a_6T_2 + a_7M + a_8F + E_5$

where: T_1 = a vector of 1s if you are in Treatment 1, zero otherwise
T_2 = a vector of 1s if you are in Treatment 2, zero otherwise
M = a vector of 1s if you are Male, zero otherwise
F = a vector of 1s if you are Female, zero otherwise

Testing the R^2 obtained from Model 4 vs. Model 5, using the F-test in equation 1, is a test of Treatment by Sex interaction.

$$F = \frac{(R^2f - R^2r)/df_1}{(1-R^2f)/df_2} \quad (1)$$

where: R^2_f = the R^2 obtained from the Full Regression model
R^2_r = the R^2 obtained from the restricted Regression model
df_1= the number of linearly independent vectors in the full model minus the number of linearly independent vectors in the restricted model
df_2= the total number of replicates (N) minus the number of linearly independent vectors in the full model

Categorical vs. Continuous Interaction (Sex and IQ)

Model 6 Full: $Y = a_0u + a_9M + a_{10}F + a_{11}M,IQ + a_{12}F,IQ + E_6$

Model 7 Restricted: $Y = a_0u + a_{13}M + a_{14}F_2 + a_{15}IQ + E_7$

where: M = a vector of 1s if you are Male, zero otherwise
F = a vector of 1s if you are Female, zero otherwise
M,IQ = a vector of 1s for Male IQs, zero otherwise
F,IQ = a vector of 1s for Female IQs, zero otherwise

Testing the R^2 obtained from Model 6 vs. Model 7, using the F-test in Equation 1, is a test of Sex by IQ interaction.

Continuous vs. Continuous Interaction (IQ and Motivation)

Model 8 Full: $Y = a_0u + a_{16}IQ + a_{17}Mot. + a_{18}IQxMot + E_8$

Model 9 Restricted: $Y = a_0u + a_{19}IQ + a_{20}Mot. + E_9$

where: IQ = a vector of continuous IQ scores
Mot. = a vector of continuous Motivation scores
IQxMot. = a vector of IQ scores x Motivation scores

Testing the R^2 obtained from Model 8 vs. Model 9, using the F-test in Equation 1, is a test of the interaction between IQ and Motivation.

Adequacy of Sample Size. The last issue addressed in this article is adequacy of the sample size. Adequate sample size**,** as discussed in Newman, Newman & Salzman (2010), clearly indicates the need to understand the importance of the underlying assumption of the N size needed for HLM when testing effects at different levels. For example, 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 simulated data suggests that for statistical power, the number of groups is more important than the number of observations. Hox (1995) and Hox and Maas (2001) reported similar findings related to the adequacy of sample size. They found that samples < 20 were insufficient at the higher levels, and if these higher-level variables were crucial to the structural model, then the N should be > 100. Obviously, the adequacy of the N size effects the power of HLM to detect interaction. However, these findings are not consistent with the position taken by Raudenbush and Bryk (2002) who believe that a Bayesian estimation approach allows the researcher to use smaller sample sizes.

There are other concerns related to determining an adequate sample size. Most researchers who are concerned with the issue of sample size know the rule of thumb, which states that 10 to 20 subjects are needed per variable for MRL. However, the rule of thumb needs to be considered as it relates to the research question. For example, a different ratio of subjects to variables is needed if one is interested in estimating Type I or Type II Error or if one wants to estimate weight stability for prediction purposes from sample to sample or sample to population, or if the researcher is interested in the models replicability (i.e., somewhat similar to cross validation) (Newman, McNeil & Fraas, 2004). Also, when using HLM models, the ratio of the number of subjects to variables is more important for the higher-level variables when estimating power. Even though these three concepts, significance; stability of weights; and replicability are related, they are different and require different numbers of subjects to adequately estimate them (Newman et al., 2004). Thus, these issues need to be considered as they relate to the research question of interest.

Comparison between Regression with Person Vectors and HLM Results

There are different computer packages for HLM on the market. For the purpose of this article, HLM 7.0 was used to perform the comparisons. Two sets of HLM models were created to test the effect of treatment over time on pain. The first set of HLM models used a fixed interaction error term and the second set allowed the error term to vary. The model for the fixed interaction error term is presented below. As can be seen, the absence of r_{1i} in the second line on the Level-2 model indicates that the error term is fixed and this becomes more apparent when looking at the mixed model.

Level-1 Model: $PAIN_{ti} = \pi_{0i} + \pi_{1i}*(TIME_{ti}) + e_{ti}$

Level-2 Model: $\pi_{0i} = \beta_{00} + \beta_{01}*(TX_i) + r_{0i}$

$\pi_{1i} = \beta_{10} + \beta_{11}*(TX_i)$

Mixed Model: $PAIN_{ti} = \beta_{00} + \beta_{01}*TX_i + \beta_{10}*TIME_{ti} + \beta_{11}*TX_i*TIME_{ti} + r_{0i} + e_{ti}$

Table 2. Final Estimation of Fixed Effects

Fixed Effect	Coefficient	*SE*	*t*-ratio	Approx. *df*	*p*-value
For INTRCPT1, π_0					
INTRCPT2, β_{00}	9.744	0.131	74.369	48	<0.001
TX, β_{01}	-1.556	0.892	-1.746	48	0.087
For TIME slope, π_1					
INTRCPT2, β_{10}	-1.169	0.031	-37.166	248	<0.001
TX, β_{11}	0.231	0.194	1.188	248	0.236

The final results, noted in Table 2, were obtained from the11th iteration. Iterations were stopped due to small changes in the likelihood function with σ^2 = 0.84096 and τ $INTRCPT1_{,\pi 0}$ = 0.11029 and a reliability estimate of 0.440. The value of the log-likelihood function at iteration 11 = -4.170159E+002 as presented in the HLM 7.0 result tables. If one were to look at the results of the final fixed effects, both TX

(p = 0.087) and the effect of TX on the slope of time (TX * Time) (p = 0.236), it would appear that TX and the TX by Time interaction were not statistically significant.

Next, the model that allowed the error term to vary was computed. As can be seen, the r_{1i} in the second line of the Level-2 Model indicates that the error term is no longer fixed.

Level-1 Model: $PAIN_{ti} = \pi_{0i} + \pi_{1i}*(TIME_{ti}) + e_{ti}$

Level-2 Model: $\pi_{0i} = \beta_{00} + \beta_{01}*(TX_i) + r_{0i}$

$\pi_{1i} = \beta_{10} + \beta_{11}*(TX_i) + r_{1i}$

Mixed Model: $PAIN_{ti} = \beta_{00} + \beta_{01}*TX_i + \beta_{10}*TIME_{ti} + \beta_{11}*TX_i*TIME_{ti} + r_{0i} + r_{1i} + e_{ti}$

Table 3. Final Estimation of Fixed Effects

Fixed Effect	Coefficient	Standard Error	*t*-ratio	Approx. df	*p*-value
For INTRCPT1, π_0					
INTRCPT2, β_{00}	9.74	0.158	61.676	48	<0.001
TX, β_{01}	-1.55	1.089	-1.427	48	0.160
For TIME slope, π_1					
INTRCPT2, β_{10}	-1.168	0.035	-33.748	48	<0.001
TX, β_{11}	0.230	0.221	1.042	48	0.303

The final results in Table 3 were obtained from the1024th iteration. Iterations were stopped due to small changes in the likelihood function. The reliability estimate for the intercept was 0.433 and for time it was 0.217. The value of the log-likelihood function at iteration 1024 = -4.140787E+002 as presented in the HLM 7.0 result tables. When looking at the results of the final fixed effects, both TX (p = 0.160) and the effect of TX on the slope of time (TX * Time) (p = 0.303), it would appear that TX and the TX by Time interaction were still not statistically significant.

These results differ from the findings when using MLR with person vectors. As can be seen by looking at the model summary and the coefficients table, even after controlling for individual difference, the Treatment by Time interaction was statistically significant with a p = 0.001. The R^2_{change} = 0.004 indicated a small interaction effect. In this example, one needs to keep in mind that even though there is a great difference in p values, the R^2_{change} is very small.

Because there was significant interaction, the data were graphed in Figure 2 to provide a clearer understanding of the Treatment by Time interaction. As one can see, pain decreased for both the Treatment and Control groups over time. However, the Treatment group had a significantly greater decrease in pain over time compared to the Control group.

Table 4. Model Summary.

		Adjusted	Std.	Change Statistics				
R	R Square	R Square	Error of Estimate	R Square Change	F Change	df1	df2	Sig. F Change
0.956	0.914	0.896	0.714	0.914	51.7	51	248	0
0.958	0.918	0.9	0.699	0.004	11.198	1	247	0.001

a. Predictors: (Constant), Person_49, Time, Person_48, ... Person_3, Person_2, Person_1, Person_13, Person_38, TX
b. Predictors: (Constant), Person_49, Time, Person_48, ... Person_3, Person_2, Person_1, Person_13, Person_38, TX, TX_Time

Conclusions

It is obvious that one should not use HLM or MLR without carefully considering what is being reflected by the statistical models, the underlining assumption, and the robustness to violations. The importance of centering is dependent on the question of interest and the type of data that exist. The interaction question is also critically important. Are the multiplicative differences in slope being tested or

are they testing the multiplicative over main effects (i.e., traditional interaction)? The main effects can potentially reduce the error term and; therefore, increase the likelihood of finding significant interactions. This is one of the potential reasons why there were differences in the example provided. The importance of the error term is not just a mathematical concern, but it is also a logical research concern that is related to the research question and to how the model is intended to be applied. Fixed effects models are more robust to violations of assumptions. If the research question of interest can logically and meaningfully be expressed as a fixed effect, the researcher may want to use the more robust fixed effects assumption. When considering the sample size required to answer the question of interest, one has to be mindful not only of the level of the analysis in HLM, but also if the researcher is asking questions about statistical significance, stability of the regression weights, or replicability. In addition to the major conceptual and logical issues that are discussed in this article, there are mathematical concerns related to the different algorithm used by the various HLM programs to obtain their solutions. Also, the results when comparing HLM to MLR show significant differences related to the multiplicative effects verses the traditional interaction. The last thought is that the results in this study support Bickel (2007) who suggests that HLM is a sophisticated way to handle nested data, but it is similar to MLR when one includes person vectors. If the appropriate model and the correct error term are used, one should be able to obtain similar results regardless of whether the researcher decides to use HLM or MLR with person vectors. If the researcher obtains different results, then he or she has to take a closer look at the statistical models because the models are reflecting different questions.

Table 5. Model Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
Restricted	(Constant)	8.639	0.311		27.765	0.000
	TX	-2.255	0.172	-0.51	-13.08	0.000
	Time	-0.575	0.051	-0.444	-11.371	0.000
	Person_1	-0.333	0.412	-0.021	-0.809	0.419
	Person_48	0.167	0.412	0.011	0.405	0.686
	Person_49	0.5	0.412	0.032	1.214	0.226
Full	(Constant)	8.306	0.321		25.893	0.000
	TX	-1.096	0.386	-0.248	-2.843	0.005
	Time	-0.408	0.07	-0.315	-5.82	0.000
	Person_1	-0.333	0.404	-0.021	-0.825	0.410
	Person_2 ...	-0.667	0.404	-0.042	-1.651	0.100
	Person_49	0.5	0.404	0.032	1.238	0.217
	TX*Time	-0.332	0.099	-0.384	-3.346	0.001

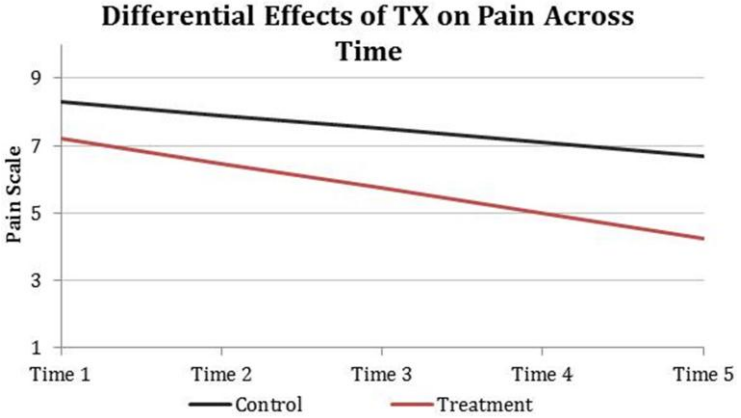

Figure 2. Treatment by time interaction.

References

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

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

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

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

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.

Kirk, R. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA:Brooks/Cole

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

Kreft, 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.

Lee. V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*, 125-141.

McCulloch, C., & Searle, S. (2001). *Generalized, linear, and mixed models*. New York Wiley & Sons.

McNeil, K., Newman, I., & Fraas, J. (2012). *Designing general linear models to test research hypotheses.* Lanham, MD: University Press of America.

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

Mundform, 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. (2002). Research practices that produce Type VI Errors. *Journal of Research in Education, 12*, 138-145.

Newman, I., McNeil, K., & Fraas, J. (2004). Two methods of estimating a study's replicability. *Mid-Western Educational Researcher, 12*, 36-40.

Newman, D., Newman, I., & Salzman, J. (2010). Comparing OLS and HLM models and the questions they answer: Potential concerns for Type VI Errors. *Multiple Linear Regression Viewpoints, 36*, 1-8.

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 from www.sarkisian.net/sc705/september6.pdf

Singer, J., & Willett, J. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* New York: Oxford University Press.

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

Tracz, S. M., 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 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
Florida Atlantic University
Email: dnewma14@fau.edu