ELSEVIER

Journal of School Psychology 48 (2010) 85 – 112

Journal of School Psychology

# A practical guide to multilevel modeling

James L. Peugh

*University of Virginia, Curry School of Education, Charlottesville, VA 22903-2495, United States*

Received 7 September 2009; accepted 8 September 2009

**Abstract**

Collecting data from students within classrooms or schools, and collecting data from students on multiple occasions over time, are two common sampling methods used in educational research that often require multilevel modeling (MLM) data analysis techniques to avoid Type-1 errors. The purpose of this article is to clarify the seven major steps involved in a multilevel analysis: (1) clarifying the research question, (2) choosing the appropriate parameter estimator, (3) assessing the need for MLM, (4) building the level-1 model, (5) building the level-2 model, (6) multilevel effect size reporting, and (7) likelihood ratio model testing. The seven steps are illustrated with both a cross-sectional and a longitudinal MLM example from the National Educational Longitudinal Study (NELS) dataset. The goal of this article is to assist applied researchers in conducting and interpreting multilevel analyses and to offer recommendations to guide the reporting of MLM analysis results.


*Keywords:* Multilevel modeling; Cross-sectional; Longitudinal; HLM; SPSS; SAS

Multilevel data often arise from many of the designs used in educational research, and analyzing multilevel data can pose unique challenges for applied researchers. Multilevel data tend to result from "nested" data structures (e.g., children nested within classrooms or schools, family members nested within families, employees nested within a business). In educational research studies, the total sample size is often a combination of students sampled from different classrooms or schools. In the multilevel analysis framework, repeated measurements in a

*E-mail address:* jp3za@Virginia.edu.
ACTION EDITOR: Craig Enders.

0022-4405/$ - see front matter © 2009 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsp.2009.09.002

longitudinal educational study are also viewed as a nested data structure, where multiple observations are nested within individuals. The problem with nested data structures is that they violate the independence assumption required by traditional statistical analyses such as ANOVA and ordinary least-squares (OLS) multiple regression. For example, the response variable scores of students in the same school are likely to be more correlated than the scores for students in different schools because they share the same environment. These independence violations tend to make multilevel modeling a necessity because traditional analysis models can produce excessive Type I errors and biased parameter estimates.

The purpose of this article is to guide applied researchers through a series of seven major steps needed to conduct multilevel modeling (MLM) analyses: (1) clarifying the research question under investigation, (2) choosing the correct parameter estimation method (i.e., full information or restricted maximum likelihood), (3) assessing whether MLM is needed, (4) building the level-1 model, (5) building the level-2 model, (6) reporting multilevel effect sizes, and (7) testing competing multilevel models using the likelihood ratio test. These seven major steps are not intended as an exhaustive list of the necessary and sufficient steps required for conducting multilevel analyses. For example, the details of additional steps in MLM analyses such as collinearity assessments and outlier diagnostics are omitted here due to space limitations. The primary goal of this article is to explain the major decision-making steps needed to enable applied researchers to conduct, interpret, and present the results of multilevel analyses. A secondary goal of this article is to provide a pedagogical summary of the numerous resources available to readers wanting additional details on MLM analyses (e.g., Clements, Bolt, Hoyt, & Kratochwill, 2007; Graves & Frohwerk, 2009; Peugh & Enders, 2005; Raudenbush & Bryk, 2002; Singer & Willett, 2003; Snijders & Bosker, 1999).

### *Analysis Example 1: Cross-Sectional MLM*

This article uses a cross-sectional model and a longitudinal model example from the National Educational Longitudinal Study (NELS; NELS: 88/2000 public use data files; National Center for Education Statistics [NCES], 2002) data set for illustration purposes; no theoretical research questions are tested and no empirical inferences should be drawn from the presented results. The NELS data set consists of various student academic achievement and school environment variables collected from $N = 12{,}144$ 8th grade students in 1988, 10th grade students in 1990, and 12 grade students in 1992 enrolled in 1476 schools (approximately 8 students per school on average) across the US. The seven major analysis steps will be illustrated first in the context of a cross-sectional example involving the prediction of science achievement scores using student socioeconomic status (SES) and each school's student-to-teacher ratio as predictor variables. As will be shown below, the primary research questions were: (a) Is multilevel modeling needed for science achievement scores?, (b) Is there a relationship between SES and student-level science achievement scores?, (c) Does the effect of SES on science achievement scores vary significantly across schools?, and (d) Is the effect of SES on science achievement scores moderated by the school's student-to-teacher ratio? Finally, the SPSS and SAS syntax commands used to conduct each cross-sectional and longitudinal example analysis are included in Appendix A.

## Clarifying the Research Question

The first step in any data analysis situation involves clarifying the research question, which is particularly important in a multilevel model. As will be shown, clarifying the research question will facilitate the analysis decisions made in subsequent steps. In general, educational researchers are often interested in research questions that focus primarily on a level-1 (e.g., student-level) variable, primarily on a level-2 (e.g., school-level) variable, or an interaction between variables. Research questions at level-1 address the effect of individual-level predictors on student-level response variables. For example, from the NELS data, a researcher may want to test a level-1 research question that addresses the effect of student SES on student science achievement. Although seemingly a straightforward analysis at the student-level, this research question still requires MLM. The science achievement variation that occurs across schools, although usually considered a nuisance in this situation, must be appropriately modeled. However, sometimes educational researchers are interested in level-2 research questions that assess the impact of school-level variables on school-level outcomes. A level-2 research question example from the NELS data might involve the effect of the student-to-teacher ratio on school-level science achievement. Further, interaction research questions examine whether the relationship between a predictor variable and a response variable is moderated by the magnitude of a third variable. From the NELS dataset, an example of an interaction research question might involve investigating whether the relationship between student SES and science achievement is moderated by the ratio of students-to-teachers in a school.

The previous three research question examples are not intended to imply an absence of predictors at an analysis level other than the level of interest. For example, researchers often want to test research questions involving the impact of a school-level variable while controlling for a student-level predictor. This is an example of another type of level-2 research question because the primary interest is the relationship at level 2. Many other types of MLM research questions could be asked, but these three cover many common research scenarios. As will be shown in steps four and five below, clarification of the multilevel research question of interest assists in choosing the correct predictor variable centering method and in building the level-1 and level-2 models prior to MLM analysis.

## Choosing an Estimation Method

MLM software packages generally give researchers the choice between two maximum likelihood (ML) estimators: full information maximum likelihood (FIML) and restricted maximum likelihood (REML). At first glance, this is a technical issue that may tempt researchers into relying on the default settings of their statistical software package. However, the choice of estimator impacts parameter estimates and nested-model test results, so researchers will benefit from making informed decisions about the two estimators.

The key difference between FIML and REML involves how MLM estimates variances. FIML effectively assumes that the MLM regression coefficients (e.g., the $\gamma_{00}$ coefficient in Equation 3 below) are known population parameters, so no degrees of freedom are allocated to these parameters during estimation. Conceptually, FIML estimates variances in

a manner akin to using population formulas with only $N$ in the denominator. Consequently, the resulting variance estimates are underestimated, although the magnitude of this bias tends to be negligible in moderate to large sample sizes. Conversely, REML treats the regression coefficients as unknown quantities to be estimated based on sample data and subtracts the needed degrees of freedom when computing variance estimates. Conceptually, REML estimates variances in a manner akin to using sample data formulas with $N-1$ in the denominator (see Singer & Willett, 2003). Further, as shown elsewhere (Raudenbush & Bryk, 2002; Singer & Willett, 2003), FIML can produce negatively biased level-1 and level-2 variance estimates under small sample size conditions, especially as the number of estimated parameters increases. While there may seem to be advantages to using REML (e.g., more accurate variance estimates in smaller sample sizes), there are important reasons to consider FIML.

ML estimation uses a log-likelihood ($\log L$) value to quantify the probability that the model being estimated produced the sample data. As shown below, multiplying the $\log L$ by $-2$ yields a value called a deviance that can be used to compare the relative fit of two competing models. Further, the meaning of deviance values differs between FIML and REML. Deviance values under FIML show how well the regression coefficients and the variance estimates fit the sample data, whereas deviance values using REML indicate only how well the variance estimates fit the data. Researchers often use likelihood ratio tests to compare MLMs that differ both in regression coefficient and variance component estimates, in which case FIML estimation is needed. In contrast, REML only allows for tests of models that differ in their variances. It is important to note that MLM statistical analysis software packages such as SPSS and SAS specify REML estimation by default. Researchers wanting to test models that differ in their regression coefficients must override this default and specify FIML estimation. An example of such a likelihood ratio test is shown below. However, it is worth noting that a review of MLM Monte Carlo simulation studies comparing REML and FIML found no clear advantage for either parameter estimation method (Kreft & de Leeuw, 1998), indicating that the choice of parameter estimator should be driven by the research question under investigation.

## Is Multilevel Modeling Needed?

Prior to the analysis of any nested dataset, the question of whether multilevel modeling is needed is a prudent one. Nested datasets do not automatically require multilevel modeling. If there is no variation in response variable scores across level-2 units (e.g., schools), the data can be analyzed using OLS multiple regression. So the question of whether MLM is needed becomes, "How much response variable variation is present at level-2?" Answering this question involves the calculation of the intraclass correlation (ICC) and the design effect statistics. Using student science achievement scores from the NELS data subset as an example, researchers can reasonably expect science achievement scores to vary across students within a school due to individual differences in ability and motivation, among other possibilities. However, science achievement scores from all students within each school can be averaged to produce a mean science achievement score for each school (Paccagnella, 2006). If there is variation in the mean science achievement levels across schools, multilevel modeling is needed to separately estimate the science achievement

variance that occurs both across students and across schools. Traditional multiple regression techniques are only designed to model response variable variance at a single unit of analysis (i.e., students or schools, but not both).

The ICC can be defined both as the proportion of science achievement score variation that occurs across schools (i.e., level-2 units) and as the expected correlation between the science achievement scores of two students (i.e., level-1 units) from the same school. Conceptually, the ICC is similar to the $R^2$ effect size from regression and the eta-squared effect size from ANOVA; it is the proportion of student science achievement score variance that can be explained by mean science achievement differences across schools. Further, the ICC provides an answer to whether MLM is needed. Specifically, an ICC value of zero indicates: (a) no mean science achievement score variation across schools (i.e., level-2), (b) all science achievement score variation occurs across students (i.e., level-1), and (c) traditional analysis techniques such as ANOVA and regression can be used to analyze the student data. However, as the ICC value increases, the proportion of science achievement score variation that occurs across schools increases, resulting in violations of the independence assumption.

A multilevel model that can partition the total science achievement score variation into its "variation across students" and "variation across schools" component parts is needed to determine if mean science achievement scores vary notably across schools (i.e., ICC > 0). The most basic MLM, an unconditional means (i.e., random effect ANOVA) model, can be shown by the following equations (Hox, 2002; Raudenbush & Bryk, 2002):

$$Level-1: Y_{ij} = \beta_{0j} + r_{ij} \quad (1)$$

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

In Eq. (1), the science achievement score of student $i$ in school $j$ ($Y_{ij}$) can be modeled as function of the mean science achievement score for school $j$ ($\beta_{0j}$) plus a residual term that reflects individual student differences around the mean of school $j$ ($r_{ij}$). In Eq. (2), the science achievement mean for school $j$ ($\beta_{0j}$) is modeled as a function of a grand-mean science achievement score ($\gamma_{00}$) plus a school-specific deviation from the grand mean ($u_{0j}$). Substituting Eq. (2) into Eq. (1) (i.e., allowing $\gamma_{00}+u_{0j}$ to replace $\beta_{0j}$ in the level-1 equation) yields the combined unconditional MLM equation below.

$$Combined: Y_{ij} = \gamma_{00} + u_{0j} + r_{ij}. \quad (3)$$

The combined model notation in Eq. (3) is important because it guides researchers in writing the syntax commands (see Appendix A) needed to estimate MLMs in SPSS and SAS (e.g., see Peugh & Enders, 2005; Singer, 1998). Other software packages (e.g., HLM) use the level-1 and level-2 equations for model specification.

The unconditional model in Eq. (3) is often referred to as a "random effects ANOVA" MLM because it partitions science achievement variability into within-group (i.e., level-1, $r_{ij}$) and between-group (i.e., level-2, $u_{0j}$) components. Note that the MLM estimates the variances of the level-1 and level-2 residuals, not the actual residuals themselves; the variance of $r_{ij}$ is denoted by $\sigma^2$, and the variance of $u_{0j}$ is denoted by $\tau_{00}$. The variation in science achievement scores at level-1 (i.e., $\sigma^2$) is the average variance of individual

students' scores within schools and is analogous to $MS_{within}$ in an ANOVA analysis. The variation in science achievement scores at level-2 (i.e., $\tau_{00}$) quantifies the variation in mean science achievement scores across schools and is analogous to $MS_{between}$ in ANOVA. The primary difference between the MLM and ANOVA is that the MLM framework views the "groups" (e.g., schools) as a random sample from a larger population of groups (i.e., a random factor), whereas ANOVA views groups as being qualitatively different (i.e., a fixed factor).

The multilevel model shown in Eq. (3) was estimated and results are shown in the second column of Table 1. The MLM shown in Eq. (3) produced three noteworthy results. First, a significant non-zero grand-mean science achievement score was observed, $\gamma_{00}=18.90$, $p<.01$. Second, the level-1 variance estimate showed significant science achievement score variation across students within a school, $\sigma^2=18.67$, $p<.01$. More important to the question of whether MLM is needed, the unconditional model results also showed significant variance in the science achievement means across schools, $\tau_{00}=4.18$, $p<.01$. Substituting these variance estimates into the following ICC equation:

$$ICC = \tau_{00}/(\tau_{00}+\sigma^2). \tag{4}$$

showed that 18% (ICC=4.18/[4.18+18.67]=.18) of the science achievement variance occurred across schools. This ICC value is consistent with research (Muthén, 1991, 1994; Muthén & Satorra, 1989; Spybrook, Raudenbush, Liu, Congdon, & Martinez, 2008) that has shown ICC values between .05 and .20 to be common in cross-sectional MLM applications in social research studies. However, a non-zero ICC estimate alone does not necessarily indicate the need for multilevel analyses.

Table 1
Model summaries: cross-sectional examples.

| Parameters | Unconditional | Level-1: fixed | Level-1: random | Interaction |
|---|---|---|---|---|
| *Regression coefficients (fixed effects)* | | | | |
| Intercept ($\gamma_{00}$) | 18.90 (.07)** | 18.89 (.07)** | 18.89 (.07)** | 18.90 (.07)** |
| Student SES ($\gamma_{10}$) | – | 2.00 (.06)** | 2.00 (.07)** | 2.00 (.07)** |
| Student-to-Teacher Ratio ($\gamma_{01}$) | – | – | – | −.10 (.01)** |
| Interaction ($\gamma_{11}$) | – | – | – | −.04 (.02)** |
| *Variance components (random effects)* | | | | |
| Residual ($\sigma^2$) | 18.67 (.25)** | 17.16 (.23)** | 16.97 (.24)** | 16.97 (.24)** |
| Intercept ($\tau_{00}$) | 4.18 (.28)** | 4.41 (.28)** | 4.45 (.28)** | 4.15 (.27)** |
| Slope ($\tau_{11}$) | – | – | .54 (.21)** | .49 (.21)** |
| Covariance ($\tau_{01}$) | – | – | .68 (.19)** | .59 (.18)** |
| *Model summary* | | | | |
| Deviance statistic | 71,308.01 | 70,394.40 | 70,374.06 | 70,310.45 |
| Number of estimated parameters | 3 | 4 | 6 | 8 |

Parameter estimate standard errors listed in parentheses.
** $p<.01$.

The design effect quantifies the effect of independence violations on standard error estimates and is an estimate of the multiplier that needs to be applied to standard errors to correct for the negative bias that results from nested data. The design effect is computed by:

$$Design\ \ Effect = 1 + (n_c - 1)ICC. \tag{5}$$

Eq. (5) shows that the design effect is influenced by the ICC and the number of students per school (i.e., $n_c$). Although the average number of students per school in the NELS dataset example was relatively small ($n_c = 12{,}144 / 1476 = 8.23$), the ICC (.18) and the design effect ($1 + [8.23\text{-}1] \cdot 18 = 2.30$) both indicate the need for multilevel modeling of science achievement data. Note that some researchers believe that design effect estimates greater than 2.0 indicate a need for MLM (cf., Muthén, 1991, 1994; Muthén & Satorra, 1989, 1995).

## Building the Level-1 Model

Recall that, in the NELS data analysis example, student SES was the level-1 (i.e., student-level) predictor of science achievement scores. Results for the unconditional model (Eq. (3)) showed significant level-1 variation in NELS science achievement scores. One or more student-level predictors could be added to the level-1 model to explain this variation. However, two additional questions influence the specification of the level-1 model. First, one question that applied researchers face involves how level-1 predictors should be centered. Briefly, predictor variables in social research are often measured on an interval scale, meaning a score of zero usually has no substantive meaning. Centering involves re-scaling a predictor variable so that a value of zero can be interpreted meaningfully. A second question is whether the relationship between the predictor and the outcome is the same for all schools, or whether the relationship varies in magnitude across schools (e.g., Is the impact of SES on science achievement the same across schools, or does this regression coefficient vary from school to school?). If the regression coefficient does vary across schools, then a so-called "random effect" needs to be added to the model to account for this additional source of variation.

Two forms of centering are possible with level-1 predictors: grand-mean centering and group-mean centering (Leaving the variables uncentered is also an option, but there are few situations that warrant this approach.). In grand-mean centering, the sample mean is subtracted from each student's predictor score (i.e., $X_{ij} - \bar{X}$). With group-mean centering, the predictor mean for the school that the student attends is subtracted from the predictor scores for each student within that school (i.e., $X_{ij} - \bar{X}_j$). The nature of the research question is crucial to determining the correct centering method. For example, as shown elsewhere (Enders & Tofighi, 2007; Hofmann & Gavin, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999; Wu & Wooldridge, 2005), if the research question bears on the predictor–response variable relationship at level-1 (i.e., a level-1 research question) or if the level-1 predictor interacts with another predictor (i.e., an interaction research question), then group-mean centering gives an unbiased estimate of the relationship. Conversely, grand-mean centering level-1 predictors adjusts response variable means for the influence of the predictor in a manner similar to analysis of covariance (ANCOVA), but it also results in level-1 slope estimates that are an uninterpretable mix of the level-1 and level-2

relationships (Enders & Tofighi, 2007; Raudenbush & Bryk, 2002). Consequently, grand-mean centering is usually best suited for research questions that involve a level-2 variable (i.e., level-2 questions). In addition, grand-mean centering is the only method that can be applied to level-2 predictors. Other resources are available that give a more detailed treatment of the centering issue (e.g., see Enders & Tofighi, 2007; Hofmann & Gavin, 1998; Kreft, de Leeuw, & Aiken, 1995; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999; Wu & Wooldridge, 2005). As a final note, the centering method at level-1 is unrelated to the variable's scale and is solely determined by the research question. For example, group- or grand-mean centering should be applied to both continuous and categorical (e.g., dummy variables) in the level-1 model (Enders & Tofighi, 2007). Leaving code variables uncentered in the level-1 model can lead to undesirable interpretations of the model parameters.

From the NELS dataset example, the level-1 predictor indicating student SES was group-mean centered because the ultimate research question involves an interaction between the level-1 predictor (i.e., student SES) and the level-2 predictor (i.e., student-to-teacher ratio). Group-mean centered SES was added to Eq. (1) to explain level-1 science achievement variation as follows:

$$Level-1: Y_{ij} = \beta_{0j} + \beta_{1j}(SES_{ij} - \overline{SES_j}) + r_{ij} \tag{6}$$

Further, adding student SES to the model raises the additional question of whether the impact of SES on student science achievement scores should be estimated as a fixed effect only (i.e., the impact of student SES on science achievement does not vary across schools), or as a fixed effect with a random effect added (i.e., the impact of SES on science achievement varies across schools). If the impact of student SES on science achievement is estimated as a fixed effect only, the level-2 intercept and slope equations for the level-1 model in Eq. (6) are:

$$Level-2: \beta_{0j} = \gamma_{00} + u_{0j} \tag{7}$$

$$Level-2: \beta_{1j} = \gamma_{10}. \tag{8}$$

Eq. (7) is identical to Eq. (2) above in its interpretation (e.g., $\gamma_{00}$ is the grand mean, and $u_{0j}$ is a residual that allows the achievement means to vary across schools). Eq. (8) illustrates the definition of a fixed effect: the impact of SES on science achievement across each school ($\beta_{1j}$) is captured by a single estimate (i.e., a fixed effect) that expresses the average effect of SES on science achievement across all schools (i.e., $\gamma_{10}$). Substituting Eqs. (7) and (8) into Eq. (6) yields the combined regression model:

$$Combined: Y_{ij} = \gamma_{00} + \gamma_{10}(SES_{ij} - \overline{SES_j}) + u_{0j} + r_{ij}. \tag{9}$$

The MLM shown in Eq. (9) was estimated and results are shown in the third column of Table 1. Results again showed a significantly non-zero mean science achievement score ($\gamma_{00}=18.89$, $p<.01$) and a significant average SES-science achievement slope ($\gamma_{10}=2.00$, $p<.01$). The regression slope indicates that science achievement increased as SES increased, such that a one-point increase in SES is associated with a two-point increase in

achievement, on average. Again, because SES scores were centered at the group mean, the $\gamma_{10}$ coefficient is an unbiased estimate of the level-1 (i.e., student-level) relationship. The variance estimates also showed significant variation in science achievement scores across students ($\sigma^2 = 17.16$, $p < .01$) and across schools ($\tau_{00} = 4.41$, $p < .01$).

Eq. (8) implies that the average impact of SES on science achievement does not differ significantly across schools (i.e., Eq. (8) lacks a $u_{1j}$ term). However, the possibility exists that some schools may show more or less than a two-point increase ($\gamma_{10} = 2.00$) in science achievement for every unit increase in SES. Said differently, school-specific deviations from the average slope coefficient may need to be added to the model. If the impact of SES on science achievement does vary significantly across schools, a variance component (i.e., random effect) would need to be added to the level-2 slope equation to model this variation, as follows:

$$Level - 2 : \beta_{1j} = \gamma_{10} + u_{1j} \tag{10}$$

The $u_{1j}$ residual term is often referred to as a random effect because it indicates that the impact of SES on science achievement is allowed to vary randomly across schools. Consistent with the previous models, the MLM does not estimate the $u_{1j}$ residuals, but the variance of these residuals, $\tau_{11}$. Substituting Eqs. (7) and (10) into Eq. (6) yields the combined MLM:

$$Combined : Y_{ij} = \gamma_{00} + \gamma_{10}(SES_{ij} - \overline{SES_j}) + u_{0j} + u_{1j}(SES_{ij} - \overline{SES_j}) + r_{ij}. \tag{11}$$

The MLM shown in Eq. (11) was estimated next; results in the fourth column of Table 1 showed identical regression coefficients ($\gamma_{00} = 18.89$ and $\gamma_{10} = 2.00$; $ps < .01$) and very similar level-1 residual variance ($\sigma^2 = 16.97$, $p < .01$) and level-2 intercept variance estimates ($\tau_{00} = 4.45$, $p < .01$). Results also showed the effect of SES on science achievement did vary significantly across schools ($\tau_{11} = .54$, $p < .01$). Interpreting the intercept variance ($\tau_{00} = 4.45$) and the SES-science achievement slope variance ($\tau_{11} = .54$) estimates is made easier by converting these variances to standard deviations (i.e., $\sqrt{4.45} = 2.11$; $\sqrt{.54} = .73$) because multilevel modeling assumes that school-specific intercept (i.e., $u_{0j}$) and slope (i.e., $u_{1j}$) deviations from their respective regression coefficient estimates ($\gamma_{00}$ and $\gamma_{10}$) are normally distributed. This means that 95% of plausible intercept and slope values fall between ±1.96 standard deviation units from the regression coefficient estimate. For example, 95% of schools have mean science achievement scores between (18.89 ± 1.96[2.11]) 14.75 and 23.03 and SES-science achievement slope values between (2.00 ± 1.96[.73]) .57 and 3.43.

At first glance, Eq. (11) seems to suggest that a model that allowed the effect of SES on science achievement to vary randomly across schools would differ from the model in Eq. (9) that did not allow the effect of SES to vary by only one estimated parameter (i.e., the variance of $u_{1j}$, denoted as $\tau_{11}$). However, a model that allows the effect of SES to vary randomly also includes a somewhat "hidden" parameter estimate, the covariance between intercepts and slopes, denoted as $\tau_{01}$. Consequently, a model that allows the effect of SES to vary across schools differs from a model that does not by two estimated parameters, $\tau_{11}$ and $\tau_{01}$. Results for the MLM indicated by Eq. (11) showed a significant intercept–slope

covariance ($\tau_{01} = .68$, $p < .01$). The positive covariance suggests that schools with a higher (i.e., more positive) SES-achievement slope tend to have higher achievement means.

The statistically significant SES slope variance and intercept–slope covariance estimates suggest that a model that allows the effect of SES to vary across schools is the better fitting model. As will be shown later, a nested-model likelihood ratio test provides a more accurate test of the variance components than the Wald $z$ tests that were used to assess the significance of the variance estimates. Further, two additional notes of caution warrant mention.

First, although it seems intuitively appealing to estimate MLMs with random effects for each level-1 predictor, doing so often leads to decreased statistical power and parameter estimation errors (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003). Briefly, parameter estimation errors in MLM typically take one of two forms: non-positive definite covariance matrix errors and non-convergence. Non-positive definite covariance matrix errors occur when the parameter estimation process produces implausible results, such as negative variance estimates or correlation estimates that exceed ±1. Further, a model is said to have "converged" when the model log$L$ statistic changes by a negligibly small value between consecutive parameter estimation iterations. Non-convergence occurs if the maximum number of parameter estimation iterations is reached, yet a non-negligible change in the log$L$ value between the final two estimation iterations occurs, indicating that the estimation process has not yet converged. Further, both non-positive definite covariance matrix errors and non-convergence typically result from one or more of the following conditions: (a) a small sample size, (b) imbalanced data (i.e., very small numbers of students sampled from some classrooms, larger numbers of students sampled from other classrooms), and (c) model misspecification (e.g., estimating a model with excessive numbers of level-1 random effects). Finally, non-positive definite covariance matrix errors and non-convergence can often be remedied by one or more of the following: (a) increasing the default number of parameter estimation iterations, (b) model simplification (e.g., decreasing the number of level-1 random effect estimates), and (c) increasing the variance of the response variable (i.e., multiplying all response variable values by 5) and decreasing the variances of the predictor variables (i.e., dividing all predictor variable values by 5; for additional details, see Singer & Willett, 2003).

Second, the SPSS software package provides two-tailed probability values by default. These probability values should be divided in half when testing variance estimates for significance. (However, a two-tailed test is appropriate for a significance test of the intercept–slope covariance). The SAS and HLM software packages do not share this condition.

## Building the Level-2 Model

Following the specification of the Level-1 model, the next step involves adding school-level predictors of interest. Recall that NELS science achievement results from the previous analyses showed significant variation in science achievement scores across schools (i.e., intercept variance) and that the impact of SES on science achievement also varied across schools (i.e., slope variance). These results, respectively, are reflected in the current intercept and slope models at Level-2 shown in Eqs. (7) and (10). One or more school-level

predictors could be added to the Level-2 models to explain this variation. However, researchers face two additional concerns when adding predictors to the Level-2 intercept and slope models in Eqs. (7) and (10): to which equation should the level-2 predictor be added?, and how should the level-2 predictor be centered?

If no level-1 predictors are present (e.g., see Eq. (1)), level-2 predictors can only be added to the level-2 intercept equation (e.g., see Eq. (2)), by definition. However, if a level-1 predictor is present, the question facing applied researchers becomes whether the level-2 predictor should be added to the intercept equation only, or added to both the intercept and slope equations at level-2. The answer to this question depends upon whether the research question involves an interaction. If the research question does not involve an interaction, the level-2 predictor should only be added to the level-2 intercept equation to obtain the main effect estimate for the level-2 predictor. However, if the research question reflected interest in a cross-level interaction, the level-2 predictor should be added to both the level-2 intercept and level-2 slope equations to allow the main effect and interaction estimate in a manner akin to factorial ANOVA. Returning to the NELS science achievement example, the level-2 predictor of student-to-teacher ratio (*ST_Ratio*) was added to both the intercept and slope equations because the primary interest in this example involves the cross-level interaction.

$$Level-2: \beta_{0j} = \gamma_{00} + \gamma_{01}(ST\_Ratio_j - \overline{ST\_Ratio}) + u_{0j} \tag{12}$$

$$Level-2: \beta_{1j} = \gamma_{10} + \gamma_{11}(ST\_Ratio_j - \overline{ST\_Ratio}) + u_{1j} \tag{13}$$

The main effects for both student SES and student-to-teacher ratio as well as the cross-level interaction term can be seen after Eqs. (12) and (13) are substituted into Eq. (6) to yield the combined model.

$$\begin{aligned} Y_{ij} = {} & \gamma_{00} + \gamma_{01}(ST\_Ratio_j - \overline{ST\_Ratio}) + \gamma_{10}(SES_{ij} - \overline{SES_j}) \\ & + \gamma_{11}(SES_{ij} - \overline{SES_j})(ST\_Ratio_j - \overline{ST\_Ratio}) + u_{0j} + u_{1j}(SES_{ij} - \overline{SES_j}) + r_{ij} \end{aligned} \tag{14}$$

Recall that Level-1 predictors could be group-mean centered or grand-mean centered. Group-mean centering is not an option for Level-2 predictors because level-2 variables are constant for all students in a given school. This leads to a simpler general rule for centering Level-2 predictors: if the Level-2 predictor is categorical (i.e., a binary dummy variable), centering is not necessary (although there is nothing wrong with centering code variables), but if the Level-2 predictor is continuous, grand-mean centering is appropriate. Returning to the NELS data example, the student-to-teacher ratio variable in Eqs. (12)–(14) was grand-mean centered because a student-to-teacher ratio score of zero lacks substantive meaning.

Returning again to the NELS data example, the cross-level interaction model shown in Eq. (14) was estimated. Results presented in the fifth column of Table 1 showed a significantly non-zero grand-mean science achievement estimate ($\gamma_{00}$=18.90, $p$<.01) for average SES students in schools with average student-to-teacher ratios. Results for the main effect of student SES ($\gamma_{10}$=2.00, $p$<.01) showed that, on average, science achievement increased two points for every unit increase in student SES controlling for the influence of

student-to-teacher ratio. Results for the main effect of student-to-teacher ratio ($\gamma_{01} = -.10$, $p < .01$) showed that, on average, the mean science achievement score for a given school decreased .10 points for every unit increase in the student-to-teacher ratio controlling for student SES. Further, results for the cross-level interaction ($\gamma_{11} = -.04$, $p < .01$) showed that a one-unit increase in the student-to-teacher ratio reduced the SES-science achievement slope by .04 units on average (e.g., see Fairchild & McQuillin, 2010).

This significant interaction is illustrated graphically in Fig. 1 to facilitate its interpretation. The interaction graph in Fig. 1 was constructed by first algebraically rearranging the fixed effects portion of Eq. (14) as shown below.

$$Y_{ij} = [\gamma_{10} - \gamma_{11}(ST\_Ratio_j - \overline{ST\_Ratio})]SES_{ij} - \overline{SES_j} + [\gamma_{01}(ST\_Ratio_j - \overline{ST\_Ratio}) + \gamma_{00}] \quad (15)$$

Next, the regression coefficient estimates (i.e., $\gamma_{00}$, $\gamma_{01}$, $\gamma_{10}$, and $\gamma_{11}$) were substituted into Eq. (15).

$$Y_{ij} = [2.00 - .04(ST\_Ratio_j - \overline{ST\_Ratio})]SES_{ij} - \overline{SES_j} + [-.10(ST\_Ratio_j - \overline{ST\_Ratio}) + 18.90] \quad (16)$$

Substituting the mean student-to-teacher ratio value (i.e., 0) and solving Eq. (16) results in the simple slope for a school with an average student-to-teacher ratio.

$$Y_{ij} = 2.00(SES_{ij} - \overline{SES_j}) + 18.90 \quad (17)$$

Similarly, substituting the centered student-to-teacher ratio value that is one standard deviation above the mean (i.e., 4.61) and solving Eq. (16) results in the simple slope for a school with a high (i.e., one standard deviation above the mean) student-to-teacher ratio.

$$Y_{ij} = 1.82(SES_{ij} - \overline{SES_j}) + 18.44 \quad (18)$$

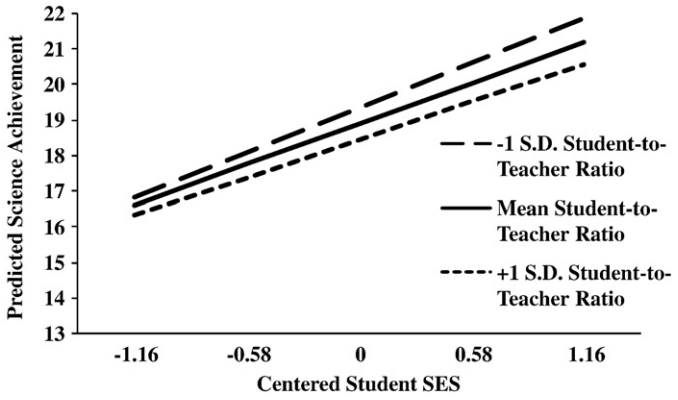


Fig. 1. Cross-sectional data example: student SES by student-to-teacher ratio interaction.

Finally, substituting the centered student-to-teacher ratio value that is one standard deviation below the mean (i.e., −4.61) and solving Eq. (16) results in the simple slope for a school with a low student-to-teacher ratio (e.g., see Aiken and West, 1991).

$$Y_{ij} = 2.18(SES_{ij} - \overline{SES_j}) + 19.36 \tag{19}$$

As shown in Fig. 1, science achievement scores increase as student SES increases, but science achievement scores increase more rapidly in schools with lower student-to-teacher ratios. However, given the sample size of the NELS example dataset ($N$=12,144), effect sizes are needed to determine whether the significant results reflect meaningful relationships or are a result of high levels of statistical power. As stated previously, these analyses are presented for illustrative purposes only and not intended to test specific research questions or draw empirical conclusions.

## Multilevel Effect Size Reporting

Effect sizes in ANOVA and multiple regression analyses, such as Cohen's $d$, eta-squared ($\eta^2$), and $R^2$, are familiar to applied researchers, and conversion formulas allow each to be placed on a similar metric to enable appropriate comparisons (see Huberty, 2002). Effect sizes in MLM analyses are not as straightforward, and currently no consensus exists as to the effect sizes that are most appropriate. The MLM effect sizes shown below are generally accepted indices (Singer & Willett, 2003, Raudenbush & Bryk, 2002) but are not comparable in the same sense as a $d$ or $\eta^2$. Interested readers can consult several sources that describe additional effect size indices that could also be used in MLM analyses (e.g., Roberts & Monaco, 2006; Snijders & Bosker, 1999).

In general, effect sizes tend to fall into two categories: global and local. Global effect sizes quantify the variance in the response variable explained by all predictor variables in an analysis model, whereas local effect sizes quantify the effect of individual variables on the response variable. In multiple regression, the global effect size $R^2$ quantifies the response variable variance explained by a model containing multiple predictors, while a squared semi-partial correlation coefficient quantifies the response variable variance accounted for by a single predictor variable, holding the influence of additional predictor variables constant. As shown below, similar global and local effect size statistics can be computed for MLMs.

One way to compute the global effect size statistic $R^2$ in multiple regression is to compute the predicted score for each participant in the sample, obtain the correlation between the observed and predicted scores, and square that correlation. Although MLM response variable variance is partitioned into Level-1 and Level-2 components, a pseudo-$R^2$ (e.g., see Singer & Willett, 2003) global effect size statistic can be computed in a similar manner. Returning to the NELS data example, computing a pseudo-$R^2$ statistic involved using the regression coefficients in Eq. (16) to obtain predicted science achievement scores ($\hat{Y}_{ij}$) for each participant in the sample, as follows.

$$\begin{aligned}\hat{Y}_{ij} = {} & 18.90 - .10(ST\_Ratio_j - \overline{ST\_Ratio}) + 2.00(SES_{ij} - \overline{SES_j}) \\ & - .04(SES_{ij} - \overline{SES_j})(ST\_Ratio_j - \overline{ST\_Ratio}).\end{aligned} \tag{20}$$

The correlation between the observed and the predicted science achievement scores (i.e., $Y_{ij}$ and $\hat{Y}_{ij}$, respectively) was $r = .26$, and squaring this value suggests that ($[.26]^2 = .07$) 7% of the variation in science achievement scores can be explained by student SES, the student-to-teacher ratio, and the interaction between these two variables.

The proportional reduction in variance statistic is a local effect size estimate that can be used in MLM analyses (Raudenbush & Bryk, 2002; Singer & Willett, 2003). Specifically, this variance reduction statistic can be computed based on the following general equation:

$$PRV = (var_{\text{NoPredictor}} - var_{\text{Predictor}}) / var_{\text{NoPredictor}}, \tag{21}$$

where PRV is the proportion reduction in variance, and "var" can represent the level-1 variance, level-2 intercept variance, or the level-2 slope variance. The "NoPredictor" subscript represents the variance estimate from the model prior to adding a predictor (e.g., the level-1 variance estimate prior to adding student SES as a predictor), and the "Predictor" subscript represents the corresponding variance from a model that contains a predictor variable (e.g., the level-1 variance estimate after adding student SES as a predictor). Returning to the NELS dataset example, the proportional reduction in Level-1 residual variance that resulted from adding student SES can be computed from the Level-1 residual variance estimates from the unconditional model ($\sigma^2 = 18.67$) and the model that includes SES ($\sigma^2 = 16.97$). Substituting these values into Eq. (16) showed that Level-1 residual variance decreased by 9% (i.e., $[18.67 - 16.97] / 18.67 = .09$) after adding student SES. Finally, it seems counterintuitive for a local effect size to be larger than a global effect size, but the pseudo-$R^2$ and proportion of variance reduction indices are not comparable, and local effect sizes that exceed global effect sizes are possible in multilevel analyses (e.g., see Hox, 2002; Roberts & Monaco, 2006; Snijders & Bosker, 1999).

The proportional variance reduction in level-2 intercept and slope variance that resulted from adding student-to-teacher ratio can also be computed. Specifically, the proportional reduction in intercept variance can be computed based on intercept variance estimates from the model containing student SES ($\tau_{00} = 4.45$) and the cross-level interaction model ($\tau_{00} = 4.15$). Results showed level-2 intercept variance decreased by 7% (i.e., $[4.45 - 4.15] / 4.45 = .07$) after adding the student-to-teacher ratio predictor to the level-2 intercept model. The proportional reduction in slope variance can also be computed from slope variance estimates from the student SES model ($\tau_{11} = .54$) and the cross-level interaction model ($\tau_{11} = .49$). Results showed level-2 slope variance decreased by 9% (i.e., $[.54 - .49] / .54 = .09$) after adding student-to-teacher ratio to the level-2 slope model. It is important to note that the proportion reduction in variance measures for level-2 variables should always be computed after adding level-1 variables to the model (Raudenbush & Bryk, 2002).

## Likelihood Ratio Model Testing

In multiple regression, an omnibus $F$ test is used to test whether the explained variance is statistically different from zero. An analogous omnibus test can be conducted in MLM analyses using the likelihood ratio test. Specifically, a likelihood ratio test can be used with the NELS example to compare the unconditional model containing no predictors to the

cross-level interaction model that contains student SES, student-to-teacher ratio, and the interaction between the two variables to test the efficacy of these predictors.

A likelihood ratio test is a statistical test of two nested models. Specifically, a "reduced" model is nested within a "full" model if the parameters estimated in the reduced model are a subset of the parameters estimated in the full model. For example, the unconditional model (reduced model; Eq. (3)) is nested within the cross-level interaction model (full model; Eq. (14)) because the parameters estimated in the unconditional model are a subset of the parameters estimated in the cross-level interaction model. Specifically, removing the $u_{1j}$, $\gamma_{01}$, $\gamma_{10}$, and $\gamma_{11}$ terms from the cross-level interaction model in Eq. (14) leaves the unconditional model in Eq. (3).

As stated previously, the deviance values (−2 times the log likelihood) for the two models can be used to decide which model better fits the data. Specifically, the difference in the deviance statistics (shown below):

$$(-2logL_{\text{ReducedModel}})-(-2logL_{\text{FullModel}}) = deviance_{\text{ReducedModel}}-deviance_{\text{FullModel}} \quad (22)$$

between the models is approximately chi-square distributed with degrees of freedom equal to the difference in the number of estimated parameters between the models, which in this case is five (the cross-level interaction model additionally estimates $\gamma_{01}$, $\gamma_{10}$, $\gamma_{11}$, $\tau_{01}$, and $\tau_{11}$). The difference in the deviance statistics between the models in Eqs. (3) and (14) is $([-2\log L_{\text{ReducedModel}}]-[-2\log L_{\text{FullModel}}])=71{,}308.01-70{,}310.45=997.56$. Referencing the likelihood ratio test to a chi-square distribution with five degrees of freedom yields a significant test statistic, $\chi^2$ (5)=997.56, $p<.01$, which showed that predicting science achievement with the cross-level interaction model was a significantly better fit to the data than predicting science achievement with the unconditional model. This likelihood ratio model test was an overall test that compared a model containing all predictor variables of interest to a model that contained none. The likelihood ratio test also could be used to test models that differ only in regression coefficient estimates (e.g., comparing a model that contained only the "main effects" of student SES and student-to-teacher ratio to a model that contains both main effects and an interaction term) or to test models that differ in variance estimates, as demonstrated below.

Some methodologists recommend using the likelihood ratio test to assess the significance of variance estimates because doing so tends to yield more accurate test results than the Wald $z$ statistic, particularly in small to moderate samples (Singer & Willett, 2003). Returning to the question of whether student SES should vary across schools (see Eq. (8)) allows an additional illustration of the likelihood ratio test. Recall that, on the basis of a statistically significant Wald $z$ test of the slope variance (i.e., $\tau_{11}$) estimate, the effect of student SES was allowed to vary across schools. Specifically, when compared to a model that estimates a constant influence of student SES (i.e., Eq. (9)), a model that allows student SES to vary across schools (i.e., Eq. (11)) includes two additional parameters: the slope variance and the intercept–slope covariance. These two models are nested because the parameters of one model are a subset of the parameters from the other model (i.e., eliminating the $u_{1j}$ term from Eq. (11) gives the model in Eq. (9)). The difference in the deviance statistics between the models in Eqs. (9) and (11) is $([-2\log L_{\text{ReducedModel}}]-[-2\log L_{\text{FullModel}}])=70394.40-70374.06=20.34$, which is distributed as a chi-square

with two degrees of freedom (i.e., the restricted model has two fewer parameter estimates, $\tau_{11}$ and $\tau_{01}$). The likelihood ratio test was significant, $\chi^2$ (2)=20.34, $p<.01$, which showed that a model allowing the effect of student SES to vary across schools fit the data significantly better than a model that assumed a constant regression coefficient across all schools. Recall that previously the effect of student SES was allowed to vary across schools on the basis of a statistically significant Wald $z$ test of the slope variance (i.e., $\tau_{11}$) estimate. This was appropriate because, at large sample sizes such as the NELS dataset example ($N$=12,144), the Wald $z$ test of variance estimates and the likelihood ratio test are equivalent. However, it is worth noting that at smaller sample sizes that are more common in applied research studies, the likelihood ratio test is the more appropriate test of nested models.

Finally, a few words of caution regarding likelihood ratio tests warrant mention. Results from a likelihood ratio test will be inaccurate, even if two models are nested, if: (a) an incorrect parameter estimator was used, (b) if one or more predictor variables have missing data, and (c) if the variables are not normally distributed. As stated previously, the results of a likelihood ratio test will be incorrect if the REML estimator was used to estimate two nested models that differ only in their regression coefficient estimates. Further, if one or more predictor variables have missing data, listwise deletion of missing data (the default missing data handling mechanism in MLM) will result in models that differ in their sample sizes. Likelihood ratio test results will be inaccurate if the nested models differ in their sample sizes (see the Baraldi & Enders, 2010 paper in this issue for additional details on missing data handling). Finally, likelihood ratio tests will also yield inaccurate results if model variables are non-normally distributed.

### *Analysis Example 2: Longitudinal MLM*

Longitudinal designs are another common example of nested data where repeated measurements (i.e., the level-1 units) are nested within individuals (i.e., the level-2 units). Longitudinal MLMs (often referred to as growth curve models) are exceedingly common in the literature, so it is important to consider these models. Accordingly, the seven MLM analysis steps shown previously will also be demonstrated in the context of a longitudinal example. Specifically, another NELS data set will be used to describe changes in reading achievement scores over time (i.e., 1988, 1990, and 1992) using gender as a predictor variable of individual growth. The primary research questions for these analyses were: (a) What average growth trajectory best describes the rate of reading achievement change over time for all students? (b) Is there significant variability across students in rates of reading achievement score changes over time? and (c) Does gender explain the variation in students' rates of reading achievement score changes? Although some authors (e.g., Ferron, 1997; Raudenbush & Bryk, 2002; Singer & Willett, 2003) use a different notational scheme for longitudinal MLM analyses, the notation from the previous section will be used to maintain consistency.

The NELS reading achievement scores are appropriate for longitudinal modeling because the scores were equated using item response theory, so the scores are expressed on the same measurement scale. It is worth mentioning that other common response variables used in educational research may not be appropriate for longitudinal MLM analysis. For example,

norm-referenced scales such as grade-equivalent scores and percentile ranks are not equal-interval scales because different amounts of absolute change in the response variable are represented by equivalent score changes on these variable scales. As demonstrated elsewhere, (e.g., see Seltzer, Frank, & Bryk, 1994), using longitudinal MLM with variables that are not measured on an interval scale can result in misleading analysis results and erroneous conclusions. The Montague, Penfield, Enders, and Huang (2010) paper in this issue addresses a similar problem with using CBM scores in a longitudinal analysis.

Finally, the database structure required for longitudinal growth modeling differs notably from the data file structure used for repeated-measures ANOVA. Typically, the data from each participant is contained in a single row, and each repeated measurement is contained in a separate column. Longitudinal growth modeling analyses in SPSS and SAS require a "stacked" database structure in which each participant has multiple rows of data, one for each longitudinal measurement occasion, and the repeated measurement response variable data are "stacked" in a single column. Further, a stacked database structure requires a unique identification number for each participant because repeated measurements are nested within each individual. Finally, a stacked database structure also requires a variable that indicates the timing of the repeated assessments. As shown below, this variable will be used in building the level-1 model. Readers interested in more information on the file format for a longitudinal MLM analysis can consult Peugh and Enders (2005) and Singer and Willett (2003).

## Is Multilevel Modeling Needed?

The question of whether MLM is needed in longitudinal data scenarios is more straightforward because students are the level-2 analysis unit, and reading achievement scores can intuitively be expected to vary significantly across students. However, to confirm this, an unconditional means (i.e., random effect ANOVA) model can be estimated to compute ICC and design effect statistics (see Eqs. (4) and (5)).

$$Level-1: Y_{ti} = \beta_{0i} + r_{ti} \tag{23}$$

$$Level-2: \beta_{0i} = \gamma_{00} + u_{0i}. \tag{24}$$

$$Combined: Y_{ti} = \gamma_{00} + u_{0i} + r_{ti}. \tag{25}$$

As stated previously, the MLM notational scheme from the previous section will be used, but the subscript notation now differs slightly from the cross-sectional examples above. The subscript $t$ now refers to repeated response variable observations (level-1 units) collected from $i$ students (level-2 units) over time. The model shown in Eqs. (23)–(25) is referred to as an unconditional means model because the reading achievement score for student $i$ at time $t$ ($Y_{ti}$) is modeled in Eq. (23) as a function of each student's mean reading achievement score ($\beta_{0i}$) plus a time-specific residual term that reflects the differences between each student's observed and predicted reading achievement scores ($r_{ti}$). In Eq. (24), each student's mean reading achievement score ($\beta_{0i}$) is modeled as a function of a grand-mean reading achievement score for all students ($\gamma_{00}$) plus a term ($u_{0i}$) that reflects deviations in an individual's reading mean around the grand mean.

The unconditional means model shown in Eq. (25) was estimated using the NELS reading achievement data described above and results are presented in the second column of Table 2. Results showed a significant grand-mean reading achievement score, $\gamma_{00}=30.07$, $p<.01$. Results also showed that students' mean reading scores (i.e., the average score across all three assessments) significantly varied around the grand mean, $\tau_{00}=70.69$, $p<.01$, as well as significant differences between each student's observed and predicted reading achievement scores over time, $\sigma^2=24.43$, $p<.01$. In the longitudinal context, the level-1 residual variance captures within-person variation in the scores (i.e., the variability of an individual's score around his or her mean), and the level-2 variance reflects individual differences between students. As shown previously, the level-2 variance estimate can be converted to a standard deviation (i.e., $\sqrt{70.69}=8.41$) to facilitate its interpretation. Specifically, assuming that the residuals are normally distributed, 95% of students had mean reading achievement scores between 13.59 and 46.55 (i.e., ±1.96 deviations from the mean; $30.07\pm1.96[8.41]$). Further, ICC calculations showed that 74% (i.e., $70.69/[70.69+24.43]=.74$) of reading achievement variation occurred across students. This value is consistent with research (Spybrook, Raudenbush, Liu, Congdon, & Martinez, 2008) that has shown ICC values that exceed .40 to be common in longitudinal social research studies. The ICC value and the design effect ($DE=1+[8.23-1].74=6.35$) both indicated the need for multilevel modeling.

## Building the Level-1 Model

Although not readily apparent, the unconditional means model describes the change in each student's reading achievement scores over time as a flat line with a slope of zero

Table 2
Model summaries: longitudinal examples.

| Parameters | Unconditional | Level-1 | Level-2: main effects | Level-2: interaction |
|---|---|---|---|---|
| *Regression coefficients (fixed effects)* | | | | |
| Intercept ($\gamma_{00}$) | 30.07 (.08)** | 27.25 (.08)** | 26.50 (.11)** | 26.52 (.11)** |
| Time ($\gamma_{10}$) | – | 1.41 (.01)** | 1.41 (.01)** | 1.38 (.02)** |
| Gender ($\gamma_{01}$) | – | – | 1.44 (.16)** | 1.40 (.16)** |
| Interaction ($\gamma_{11}$) | – | – | – | .07 (.03)* |
| *Variance components (random effects)* | | | | |
| Residual ($\sigma^2$) | 24.43 (.22)** | 13.63 (.17)** | 13.63 (.17)** | 13.63 (.17)** |
| Intercept ($\tau_{00}$) | 70.69 (1.01)** | 63.97 (.98)** | 63.49 (.97)** | 63.49 (.97)** |
| Slope ($\tau_{11}$) | – | .71 (.04)** | .71 (.04)** | .71 (.04)** |
| Covariance ($\tau_{01}$) | – | 1.87 (.13)** | 1.84 (.13)** | 1.85 (.13)** |
| *Model summary* | | | | |
| Deviance statistic | 247,383.31 | 236,739.05 | 236,654.48 | 236,648.97 |
| Number of estimated parameters | 3 | 6 | 7 | 8 |

Note: Parameter estimate standard errors listed in parentheses.
* $p<.05$.
** $p<.01$.

located at each student's mean reading achievement score. Adding a level-1 'time' predictor to the model allows the changes in each student's reading achievement scores over time to be modeled with a straight line with a non-zero slope.

$$Level-1: Y_{ti} = \beta_{0i} + \beta_{1i}(TIME_{ti}) + r_{ti} \tag{26}$$

$$Level-2: \beta_{0i} = \gamma_{00} + u_{0i} \tag{27}$$

$$Level-2: \beta_{1i} = \gamma_{10} + u_{1i} \tag{28}$$

$$Combined: Y_{ij} = \gamma_{00} + \gamma_{10}(TIME_{ti}) + u_{0i} + u_{1i}(TIME_{ti}) + r_{ti} \tag{29}$$

Contrary to the descriptions given for terms in the unconditional means model (Eqs. (23)–(25)), the terms in Eqs. (26)–(29) are now conditional on the chosen metric of time. The term initial status will now be used to further describe the intercept ($\beta_{0i}$) as the expected reading achievement score for a student at the measurement occasion when the chosen time metric equals zero. Specifically, as shown in Eq. (26), the reading achievement score of student *i* at time *t* ($Y_{ti}$) is modeled as a function of a student's reading achievement score ($\beta_{0i}$) at the intercept or at initial status (i.e., when time=0) plus a term that reflects how each student's reading achievement score changes over time ($\beta_{1i}$) plus a residual term that reflects the difference between each student's observed and predicted reading achievement scores ($r_{ti}$). As shown in Eq. (27), each student's reading achievement score at the intercept, or initial status, is modeled as a grand-mean reading achievement score at initial status ($\gamma_{00}$) plus a residual term that reflects deviations in students' initial status reading achievement scores about the grand mean ($u_{0i}$). Further, as shown in Eq. (28), each student's rate of reading achievement score change across time ($\beta_{1i}$) is modeled as grand-mean rate of reading achievement change ($\gamma_{10}$) plus a residual term that reflects individual student differences in reading achievement change about the grand mean ($u_{1i}$).

Returning to the NELS dataset example, reading achievement data were collected from 8th graders in 1988, 10th graders in 1990, and 12th graders in 1992. The most meaningful metric of time in this example analysis is grade; changes in reading achievement were modeled as a function of grade level increases. Further, just as student SES was centered prior to its inclusion into the cross-sectional level-1 MLM to facilitate its interpretation, centering the chosen metric of time provides both a location and an interpretation for the intercept (Biesanz, Deeb-Sossa, Papadakis, Bollen, & Curran, 2004; Mehta & West, 2000). Centering the metric of time in longitudinal MLM involves subtracting a constant value from the time metric to provide a location and interpretation for the intercept that informs the research question. In the example below, 'grade' was centered such that the intercept is interpreted as the average reading achievement score for students in grade 8. This choice of centering can be further illustrated by rewriting Eqs. (26)–(29) as follows:

$$Level-1: Y_{ij} = \beta_{0i} + \beta_{1i}(GRADE_{ij}-8) + r_{ij} \tag{30}$$

$$Level-2: \beta_{0i} = \gamma_{00} + u_{0i} \tag{31}$$

$$Level-2: \beta_{1i} = \gamma_{10} + u_{1i} \tag{32}$$

$$Combined: Y_{ij} = \gamma_{00} + \gamma_{10}(GRADE_{ij} - 8) + u_{0i} + u_{1i}(GRADE_{ij} - 8) + r_{ij} \tag{33}$$

In this example, grade was centered such that the intercept corresponded to the initial assessment time point (i.e., $GRADE_{ij} - 8$; each student's centered grade scores were 0, 2, and 4). The mean slope coefficient (i.e., the $\gamma_{10}$ coefficient in Eq. (32)) is the expected change in reading achievement for a one-unit (i.e., one-year) increment in time. The coding of time is somewhat arbitrary, but it has an important bearing on the interpretation of the growth model parameters. For example, choosing centered time values of 0, 1, and 2 would not affect the intercept, but it would change the $\gamma_{10}$ interpretation to reflect the average change per two-year interval (i.e., a one-unit change in time would correspond to a two-year interval). Several additional options are available for centering the metric of time (e.g., see Singer & Willett, 2003). For example, if the NELS reading achievement data had involved the assessment of an administered intervention, the time metric (grade) could be centered at the final measurement occasion (i.e., $GRADE_{ij} - 12$; each student having grade scores of −4, −2, and 0) so that the intercept could reflect the effect of that intervention at the end of the study. For the purposes of tracking longitudinal change, TIME can be a variable measured on any metric that is meaningful in a given research context (e.g., days, weeks, months, years). Many published applications of longitudinal MLMs fail to adequately describe the coding and centering of their time metric variable, despite the fact that these choices have an important bearing on the interpretation of the analysis results.

Returning to the NELS reading achievement example, the MLM shown in Eq. (33) was estimated and results are presented in the third column of Table 2. Specifically, results showed significant grand-mean reading achievement score at grade 8 ($\gamma_{00} = 27.25$, $p < .01$) that increased 1.41 points per grade level, on average ($\gamma_{10} = 1.41$, $p < .01$). Further, variance component estimates showed: (a) significant variance in observed versus predicted reading achievement within students (level-1 residual; $\sigma^2 = 13.63$, $p < .01$), (b) significant variation in reading achievement scores across students at grade 8 ($\tau_{00} = 63.97$, $p < .01$), and (c) significant slope variance ($\tau_{11} = .71$, $p < .01$) in reading achievement growth trajectories across students. Converting the intercept variance estimate to a standard deviation (i.e., $\sqrt{63.97} = 8.00$), and assuming intercept residuals to be normally distributed, suggests 95% of students' grade 8 reading achievement scores would be expected to fall between 11.57 and 42.93 (i.e., ±1.96 standard deviations from the mean intercept estimate; $27.25 \pm 1.96$ [8.00] = 11.57, 42.93). Further, converting the slope variance estimate to a standard deviation (i.e., $\sqrt{.71} = .84$), and assuming normally distributed slope residuals, suggests 95% of students would be expected to have reading achievement scores that changed between $(1.41 \pm 1.96[.84] = -.24, 3.06) - .24$ and 3.06 reading achievement score points per grade. Finally, the intercept–slope covariance estimate ($\tau_{01} = 1.87$, $p < .01$) also was significant. A positive intercept–slope covariance estimate showed that students with higher reading achievement scores at grade 8 showed greater increases in reading achievement across grades 10 and 12.

As shown in Eq. (30), the effect of linear growth is tested because the NELS reading achievement data were collected at only three assessment occasions. Additional repeated measurement occasions would be needed to test the effect of higher-order growth forms

(i.e., quadratic). For example, if additional repeated measurements were available, the effect of quadratic growth could be tested by adding a powered vector (i.e., $\beta_{2i}$ $[GRADE_{ij}-8]^2$) to the level-1 model in Eq. (30). Readers interested in specifying more complex level-1 growth forms can consult several sources for specific details (e.g., Raudenbush & Bryk, 2002; Singer & Willett, 2003; Snijders & Bosker, 1999).

## Building the Level-2 Model

The previous level-1 longitudinal model indicated significant intercept and slope variance in reading achievement growth across students. A binary level-2 predictor variable, gender (i.e., female = 1, male = 0), was added to the Level-2 model to explain intercept and slope variance in reading achievement. As shown below, gender was added to the Level-2 models uncentered because, by definition, a binary dummy variable has a meaningful zero point (although centering gender is also appropriate).

$$Level-1: Y_{ti} = \beta_{0i} + \beta_{1i}(GRADE_{ti}-8) + r_{ti} \tag{34}$$

$$Level-2: \beta_{0i} = \gamma_{00} + \gamma_{01}(GENDER_i) + u_{0i} \tag{35}$$

$$Level-2: \beta_{1i} = \gamma_{10} + \gamma_{11}(GENDER_i) + u_{1i} \tag{36}$$

$$Combined: Y_{ti} = \gamma_{00} + \gamma_{01}(GENDER_i) + \gamma_{10}(GRADE_{ti}-8) + \gamma_{11}(GENDER_i)(GRADE_{ti}-8) + u_{0i} + u_{1i}(GRADE_{ti}-8) + r_{ti} \tag{37}$$

Further, adding gender as a binary dummy predictor variable in Eqs. (35) and (36) changes the interpretations of the intercept ($\gamma_{00}$) and slope ($\gamma_{10}$) parameter estimates. Specifically, the intercept is now interpreted as the average reading achievement score for male students (i.e., the group coded zero), and the slope is now the average change in male students' reading achievement scores per grade level increase. Consistent with the use of dummy code predictors in multiple regression analyses, the $\gamma_{01}$ coefficient is the mean gender difference at 8th grade, and the cross-level interaction coefficient ($\gamma_{11}$) is the mean slope difference between males and females. Note that centering the gender variable would have produced identical estimates of $\gamma_{00}$ and $\gamma_{10}$ in the two analyses (the $\gamma_{01}$ and $\gamma_{11}$ coefficients are unaffected by centering).

The conditional longitudinal MLM shown in Eq. (37) was estimated and the results for all estimated parameters are shown in the fifth column of Table 2. Specifically, results showed a significant mean reading achievement score for male students in grade 8 (intercept; $\gamma_{00}=26.52$, $p<.01$) that significantly increased for male students as grade increased (slope; $\gamma_{10}=1.38$, $p<.01$). Results also showed female students had significantly higher reading achievement scores at grade 8 ($\gamma_{01}=1.40$, $p<.01$) and significantly larger increases in reading achievement scores per increase in grade level ($\gamma_{11}=.07$, $p=.02$) compared to male students. Variance component estimates again showed: (a) significant variance in observed versus predicted reading achievement within students (level-1 residual; $\sigma^2=13.63$, $p<.01$), (b) significant variation in reading achievement scores at grade 8 ($\tau_{00}=63.49$, $p<.01$), (c) significant slope variance ($\tau_{11}=.71$, $p<.01$) in reading

achievement growth trajectories across students, and (d) a significant intercept–slope covariance estimate ($\tau_{01}=.71$, $p<.01$).

Two additional findings from the variance components estimates warrant mention. The level-1 residual variance estimate ($\sigma^2=13.63$) is unchanged from the previous example analysis because the level-1 residual variance is unaffected by the addition of a level-2 predictor. Also, the level-2 intercept ($\tau_{00}=63.49$) and slope ($\tau_{11}=.71$) variance estimates decreased only slightly from the previous analysis example; this suggests a small effect size for gender. Had gender been a stronger predictor of reading achievement variance, the level-2 intercept and slope variance estimates would have decreased to a greater degree.

In the context of the longitudinal analysis, the cross-level interaction term represents the gender difference in growth rates (i.e., the time-by-gender interaction). The significant interaction term ($\gamma_{11}=.07$, $p=.02$) is illustrated graphically in Fig. 2. The interaction graph shown in Fig. 2 was constructed by first substituting the regression coefficient estimates (i.e., $\gamma_{00}$, $\gamma_{01}$, $\gamma_{10}$, and $\gamma_{11}$) into the fixed effects portion of Eq. (37).

$$Y_{ti} = 26.52 + 1.40(GENDER_i) + 1.38(GRADE_{ti}-8) + .07(GENDER_i)(GRADE_{ti}-8) \tag{38}$$

Substituting the values for gender (i.e., 0=male, 1=female) into Eq. (38) results in simple slope equations predicting reading achievement for females and males, respectively.

$$Females: Y_{ij} = 27.92 + 1.45(GRADE_{ij}-8) \tag{39}$$

$$Males: Y_{ij} = 26.52 + 1.38(GRADE_{ij}-8) \tag{40}$$

Substituting the centered values of grade (i.e., 0, 2, and 4) into the equations provides predicted reading achievement scores for males and females at each grade level (Aiken and West, 1991). Fig. 2 is a graph of predicted reading achievement scores (i.e., average growth trajectories) for male and female students. As shown in Fig. 2, female students have higher reading achievement scores that increase at a slightly greater rate over time compared to male student reading achievement scores.

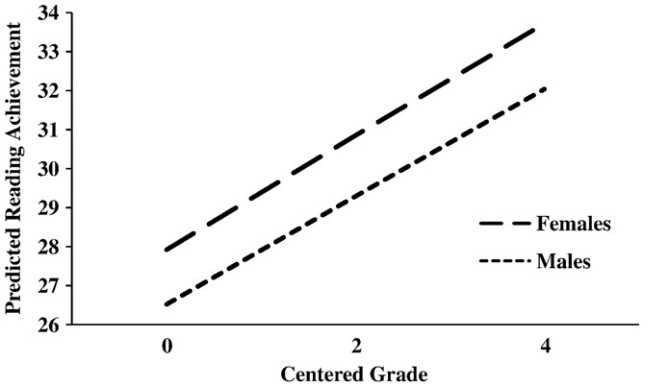


Fig. 2. Longitudinal data example: linear grade-by-gender interaction.

## Multilevel Effect Size Reporting

The global pseudo-$R^2$ effect size statistic for the longitudinal reading achievement model can be computed in the same way the pseudo-$R^2$ statistic was computed for the cross-sectional model example (see Eq. (19)). Specifically, predicted reading achievement scores ($\hat{Y}_{ti}$) were computed by solving Eq. (37) for each participant. The correlation between the observed and predicted reading achievement scores was $r = .25$; squaring this value suggests that ($[.25]^2 = .06$) 6% of the variation in reading achievement scores can be explained by linear change, gender, and the interaction between the two variables.

Local effect size statistics quantifying the proportional reduction in level-1 and level-2 variance from adding linear grade and gender, respectively, could also be computed. The proportional reduction in Level-1 residual variance that resulted from adding linear grade can be computed from the Level-1 residual variance estimates from the unconditional means model ($\sigma^2 = 24.43$) and the model that included linear grade ($\sigma^2 = 13.63$) as a level-1 predictor variable. Substituting these values into Eq. (22) showed that level-1 residual variance decreased ($[24.43 - 13.63]/24.43 = .44$) 44% after adding the linear growth term (i.e., the grade variable). Further, the proportional variance reduction in Level-2 intercept and slope variance that resulted from adding gender can also be computed. Specifically, the proportional reduction in intercept variance can be computed based on intercept variance estimates from the model containing grade level ($\tau_{00} = 63.97$) and the grade-by-gender cross-level interaction model ($\tau_{00} = 63.49$). Results showed Level-2 intercept variance decreased by ($[63.97 - 63.49]/63.97 = .01$) 1% after adding gender to the Level-2 intercept model. The proportional reduction in slope variance can also be computed from slope variance estimates from the linear grade model ($\tau_{11} = .72$) and the grade-by-gender cross-level interaction model ($\tau_{11} = .71$). Results showed Level-2 slope variance also decreased by ($[.72 - .71]/.72 = .01$) 1% after adding gender to the Level-2 slope model.

## Likelihood Ratio Model Testing

Recall that in the cross-sectional model example, two likelihood ratio model tests were conducted. The first was an overall model test that compared the cross-level interaction model to the unconditional model to assess the efficacy of the predictor variables. A similar overall model test was conducted with the NELS longitudinal reading achievement data; the cross-level gender-by-grade level interaction model (Eq. (37)) was tested against the unconditional means model (Eq. (25)). However, rather than conduct a second test of models that differ in their variance components, as was done for student SES in the cross-sectional example, a second test for the longitudinal reading achievement data will involve a test of models that differ in their regression coefficients. Specifically, a model that contains only the main effects of linear grade and gender (see column four of Table 2) will be compared to the final model (Eq. (37)) that contained both main effects and the interaction between the two predictor variables.

The overall likelihood ratio model test for the longitudinal reading achievement data involved comparing the unconditional means model (Eq. (25)) to the cross-level interaction model (Eq. (37)) to test the efficacy of linear grade and gender as predictors. The difference in the deviance statistics between the two models, $([-2\log L_{\text{ReducedModel}}] - [-2\log L_{\text{FullModel}}]) =$

247,383.31 − 236,648.97 = 10,734.34, is distributed as a chi-square with five degrees of freedom (the cross-level interaction model additionally estimates $\gamma_{01}$, $\gamma_{10}$, $\gamma_{11}$, $\tau_{01}$, and $\tau_{11}$). The likelihood ratio test was significant, $\chi^2$ (5) = 10,734.34, $p < .01$, which showed that predicting reading achievement with the cross-level interaction model was a significantly better fit to the data than predicting reading achievement with the unconditional means model.

Further, as stated previously in the cross-sectional example, the likelihood ratio model test can also be used to compare models that differ in their regression coefficient estimates. Recall that the results of the cross-level interaction model (Eq. (37)) showed a small but significant coefficient estimate for the grade-by-gender interaction term ($\gamma_{11} = .07$, $p = .02$). Recall also that local proportional reduction in intercept and slope variance statistics showed both intercept and slope variance decreases were very modest (both 1%) as a result of adding gender as a level-2 predictor. Given these results, a researcher might reasonably wonder if the interaction term ($\gamma_{11}$) could be removed from the cross-level interaction model (see Eq. (37)). The difference in the deviance statistics between the two models ($[-2\log L_{\text{ReducedModel}}] - [-2\log L_{\text{FullModel}}]$) = 236,654.48 − 236,648.97 = 5.51, is distributed as a chi-square with one degree of freedom (the main effects model does not estimate $\gamma_{11}$). The likelihood ratio test was significant, $\chi^2$ (1) = 5.51, $p < .05$, which showed that predicting reading achievement with the cross-level interaction model was a significantly better fit to the data than predicting reading achievement with a main effects model that removed the linear grade-by-gender interaction term. However, the fact that a rather large sample produced such a small chi-square statistic further underscores the point that the gender effect was relatively small in magnitude.

## Conclusion

The goal of this article was twofold. The first goal was to clarify the decisions that need to be made by applied researchers prior to MLM data analyses. The second goal was to assist applied researchers in conducting and interpreting MLM analyses and reporting the results. To further both goals, the process of conducting and interpreting MLM analyses was presented as a series of seven steps: (1) clarifying the research question under investigation, (2) choosing the correct parameter estimation method (i.e., full information or restricted maximum likelihood), (3) assessing whether MLM is needed, (4) building the level-1 model, (5) building the level-2 model, (6) reporting multilevel effect size, and (7) testing competing multilevel models using the likelihood ratio test.

As shown throughout the article, presenting the research question under investigation not only clarifies the theory being tested, but also informs MLM decision-making at each step in the analysis process. Consistent with the research question, researchers should indicate the parameter estimator used in all analyses: REML is an appropriate estimator if competing models to be tested differ only in their variance components, but FIML allows tests of competing models that differ in either their regression coefficient or variance components estimates. Following the presentation of the research question and choice of parameter estimator, researchers should first establish the need for MLM analyses by presenting the results of *ICC* and design effect statistics calculations. Further, the research question guides the specification of the level-1 and level-2 models to be analyzed and

assists researchers in choosing the appropriate centering method for level-1 and level-2 predictor variables. If the research question involves a test of competing models that are nested, researchers need to report the results of the likelihood ratio model test, placing emphasis on reporting: (a) the chi-square fit statistics and the number of parameters estimated for both models, (b) the difference in both the chi-square statistics and the number of estimated parameters between the two models, and (c) the test of the difference statistic at degrees of freedom equal to the difference in the number of estimated parameters. Once the model of best fit is found, researchers should then present both the global pseudo-$R^2$ and local proportion of variance reduction effect size statistics. To facilitate the reporting and interpreting of all MLM analysis results that could guide future investigations (not just the regression coefficient estimates that relate to the research question under investigation), all parameter estimates could be reported in tables similar to Tables 1 and 2. Further, if the research question points to a cross-level interaction, the significant interaction coefficient can be presented graphically for presentational clarity and ease of interpretation in a manner similar to the interaction graphs presented in Figs. 1 and 2.

Widely-used statistical analysis software packages such as HLM, SAS, and SPSS have made sophisticated multilevel modeling techniques readily available to applied researchers. However, no statistical analysis software package can resolve the model development and specification decisions necessary in MLM analyses (Kreft, 1995; Singer, 1998). This article is offered as an aid to guide researchers through the model development, model specification, data analysis, and results reporting decisions necessary in multilevel modeling.

## Appendix A

| SPSS Syntax: Cross Sectional | SAS Syntax: Cross Sectional |
|---|---|
| *Equation #3 Model: Unconditional Means.<br>MIXED SCIENCE_T1<br>/PRINT=SOLUTION TESTCOV<br>/METHOD=ML<br>/FIXED=INTERCEPT<br>/RANDOM=INTERCEPT \|<br>SUBJECT (SCHOOL_ID) COVTYPE(ID). | *Equation #3 Model: Unconditional Means;<br>Proc Mixed Method=ML NoCLPrint Covtest<br>NoItPrint<br>Class School_Id;<br>Model Science_T1 = /Solution;<br>Random Intercept /<br>Sub=School_Id; |
| *Equation #9 Model: Student SES Fixed.<br>MIXED SCIENCE_T1 WITH GROUP_SES<br>/PRINT=SOLUTION TESTCOV<br>/METHOD=ML<br>/FIXED=INTERCEPT GROUP_SES<br>/RANDOM=INTERCEPT \|<br>SUBJECT (SCHOOL_ID) COVTYPE(ID). | *Equation #9 Model: Student SES Fixed;<br>Proc Mixed Method=ML NoCLPrint Covtest<br>NoItPrint<br>Class School_Id;<br>Model Science_T1=Group_SES/Solution<br>ddfm=bw notest;<br>Random Intercept / Sub=School_Id Type=Un; |
| *Equation #11 Model: Student SES Random.<br>MIXED SCIENCE_T1 WITH GROUP_SES<br>/PRINT=SOLUTION TESTCOV<br>/METHOD=ML<br>/FIXED=INTERCEPT GROUP_SES | *Equation #11 Model: Student SES Random;<br>Proc Mixed Method=ML NoCLPrint Covtest<br>NoItPrint Class School_Id;<br>Model Science_T1=Group_SES/Solution<br>ddfm=bw notest; |

(continued on next page)

**Appendix A** (*continued*)

| SPSS Syntax: Cross Sectional | SAS Syntax: Cross Sectional |
|---|---|
| /RANDOM=INTERCEPT GROUP_SES \| SUBJECT (SCHOOL_ID) COVTYPE(UN). | Random Intercept Group_SES / Sub=School_Id Type=Un; |
| *Equation #14 Model: Cross-Level Interaction. MIXED SCIENCE_T1 WITH GROUP_SES GRAND_ST_RATIO /PRINT=SOLUTION TESTCOV /METHOD=ML /FIXED=INTERCEPT GROUP_SES GRAND_ST_RATIO GROUP_SES*GRAND_ST_RATIO /RANDOM=INTERCEPT GROUP_SES \| SUBJECT (SCHOOL_ID) COVTYPE(UN). | *Equation #14 Model: Cross-Level Interaction; Proc Mixed Method=ML NoCLPrint Covtest NoItPrint Class School_Id; Model Science_T1=Group_SES Grand_ST_Ratio Group_SES*Grand_ST_Ratio /Solution ddfm=bw notest; Random Intercept Group_SES / Sub=School_Id Type=Un; |
| **SPSS Syntax: Longitudinal** | **SAS Syntax: Longitudinal** |
| *Equation #25 Model: Unconditional. MIXED READING /PRINT=SOLUTION TESTCOV /METHOD=ML /FIXED=INTERCEPT /RANDOM INTERCEPT \| SUBJECT(PERSON_ID) COVTYPE(ID). | *Equation #25 Model: Unconditional; Proc Mixed Method=ML NoCLPrint Covtest NoitPrint Class Person_Id; Model Reading = /Solution; Random Intercept /subject=Person_Id; |
| *Equation #33 Model: Random Grade. MIXED READING WITH TIME /PRINT=SOLUTION TESTCOV /METHOD=ML /FIXED=INTERCEPT TIME /RANDOM INTERCEPT TIME \| SUBJECT(PERSON_ID) COVTYPE(UN). | *Equation #33 Model: Random Grade; Proc Mixed Method=ML NoCLPrint Covtest NoitPrint Class Person_Id; Model Reading=Time /Solution ddfm=bw notest; Random Intercept Time/Subject=Person_Id Type=Un; |
| *Equation #37 Model: Cross-Level Interaction. MIXED READING WITH TIME SP_RELATION /PRINT=SOLUTION TESTCOV /METHOD=ML /FIXED=INTERCEPT TIME GENDER TIME*GENDER /RANDOM INTERCEPT TIME \| SUBJECT(PERSON_ID) COVTYPE(UN). | *Equation #37 Model: Cross-Level Interaction; Proc Mixed Method=ML NoCLPrint Covtest NoitPrint Class Person_Id; Model Reading=Time Gender Time*Gender /Solution ddfm=bw notest; Random Intercept Time/Subject=Person_Id Type=Un; |
| *Modified Equation #37 Model: Main Effects. MIXED READING WITH TIME SP_RELATION /PRINT=SOLUTION TESTCOV /METHOD=ML /FIXED=INTERCEPT TIME GENDER /RANDOM INTERCEPT TIME \| SUBJECT(PERSON_ID) COVTYPE(UN). | *Modified Equation #37 Model: Main Effects; Proc Mixed Method=ML NoCLPrint Covtest NoitPrint Class Person_Id; Model Reading=Time Gender /Solution ddfm=bw notest; Random Intercept Time/Subject=Person_Id Type=Un; |

## References


Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions.* Thousand Oaks, CA: Sage.

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The role of coding time in estimating and interpreting growth curve models. *Psychological Methods*, *9*, 30−52.

Baraldi, A., & Enders, C. K. (2010). A primer on modern missing data handling methods. *Journal of School Psychology*, *48*.

Clements, M. A., Bolt, D., Hoyt, W., & Kratochwill, T. R. (2007). Using multilevel modeling to examine the effects of multitiered interventions. *Psychology in the Schools*, *44*, 503−513.

Graves, S. L., & Frohwerk, A. (2009). Multilevel modeling and school psychology: A review and practical example. *School Psychology Quarterly*, *24*, 84−94.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121−138.

Fairchild, A. J., & McQuillin, S. D. (2010). Evaluating mediation and moderation effects in school psychology: A presentation of methods and review of current practice. *Journal of School Psychology*, *48*.

Ferron, J. (1997). Moving between hierarchical model notations. *Journal of Educational and Behavioral Statistics*, *22*, 119−123.

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

Hox, J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Erlbaum.

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*, 227−240.

Kreft, I. G. G. (Ed.). (1995). *Hierarchical linear models: Problems and prospects [Special issue]Journal of Educational and Behavioral Statistics*, *vol. 20*.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage.

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.

Mehta, P. D., & West, S. G. (2000). Putting the individual back in to individual growth curves. *Psychological Methods*, *5*, 23−43.

Montague, M., Penfield, R. D., Enders, C., & Huang, J. (2010). Curriculum-based measurement of math problem solving: A methodology and rationale for establishing equivalence of scores. *Journal of School Psychology*, *48*.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338−354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376−398.

Muthén, B. O., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 87−99). San Diego: Academic Press.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology, 1995* (pp. 267−316). Oxford: Blackwell.

National Center for Education Statistics, U.S. Department of Education (2002). *NELS:88 base-year to fourth follow-up data file user's manual.*Washington, DC: U.S. Government Printing Office (NCES Publication No. 2002-323).

Paccagnella, O. (2006). Centering or not centering in multilevel models? The role of the group mean and the assessment group effects. *Evaluation Review*, *30*, 66−85.

Peugh, J. L., & Enders, C. K. (2005). Using the SPSS mixed procedure to fit cross-sectional and longitudinal multilevel models. *Educational and Psychological Measurement*, *65*, 717−741.

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

Roberts, J.K., & Monaco, J.P. (2006, April). *Effect size measures for the two-level linear multilevel model*. Paper presented at the annual conference of the American Educational Research Association [AERA], San Francisco, CA.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis*, *16*, 41−49.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, *24*, 323−355.

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Spybrook, J., Raudenbush, S. W., Liu, X. -F., Congdon, R., & Martinez, A. (2008). *Optimal Design software for multi-level and longitudinal research: Documentation for the "Optimal Design" software (Version 1.77) [Computer software].* Available at http://sitemaker.umich.edu/group-based

Wu, Y. -W. B., & Wooldridge, P. J. (2005). The impact of centering first-level predictors on individual and contextual effects in multilevel data analysis. *Nursing Research*, *54*, 212−216.