Technical Report # 1308

Hierarchical Linear Modeling (HLM): An Introduction to Key Concepts Within Cross-Sectional and Growth Modeling Frameworks

Daniel Anderson

University of Oregon

Published by

Behavioral Research and Teaching
University of Oregon • 175 Education
5262 University of Oregon • Eugene, OR 97403-5262
Phone: 541-346-3535 • Fax: 541-346-5689
http://brt.uoregon.edu

Note: Funds for this dataset were provided by the Oregon Department of Education Contract No. 8777 as part of Project OFAR (Oregon Formative Assessment Resources) Statewide Longitudinal Data System (SLDS), CFDA 84.372A Grant Program that is authorized by the Educational Technical Assistance Act of 2002.

Abstract

This manuscript provides an overview of hierarchical linear modeling (HLM), as part of a series of papers covering topics relevant to consumers of educational research. HLM is tremendously flexible, allowing researchers to specify relations across multiple "levels" of the educational system (e.g., students, classrooms, schools, etc.). The manuscript contains three chapters. In Chapter 1, the concept of HLM is introduced, as well as topics that will be covered in the paper. Chapter 2 provides a basic overview of cross-sectional HLM models, complete with an illustrated example contrasting results of an HLM model with a standard single-level regression model. The bulk of the manuscript is reserved for Chapter 3, which covers the application of HLM to modeling growth. Chapter 3, again, concludes with illustrated examples. The manuscript is concluded with an overall discussion of HLM and what was and was not covered within the manuscript.

Chapter 1: Introduction

Hierarchical linear modeling (HLM) is a powerful and flexible statistical framework for analyzing complex nested relationships. In education, for example, we may be interested in factors that affect student achievement. Broadly, we may theorize factors associated with the school (school social groups, principal leadership, school size), the teachers (effectiveness of the teacher, specific expertise of the teacher, relationship of the teacher with the student), and the students themselves (motivation, previous achievement, general intelligence). Each of these factors associated with student achievement could be conceptualized as different "levels" of nesting – students (at Level 1) are nested within classrooms (at Level 2), which are nested within schools (at Level 3) – in which each level potentially impacts student achievement. HLM allows researchers to investigate these nested relationships and either parse them out (i.e., control for higher-level factors to examine the unique effect of a specific variable) or examine the impact of variables at the higher levels (e.g., the effect of attending public versus private school on student achievement). HLM is used across a variety of disciplines to examine multilevel effects. For example, in organizational research one may investigate how employee interactions differ by the type of organization the employees belong to (e.g., corporate compared to local). In this paper, however, I will be focusing exclusively on the application of HLM to educational research.

HLM is particularly well suited for evaluating changes in student achievement through growth models applied to longitudinal data. These growth models can be used to evaluate how individuals are changing over time, and how specific variables at any level predict where the individuals begin and/or the rate at which they change. For example, we could examine data from a cohort of students as they moved from kindergarten through grade 5 in urban and rural schools. We could then test whether the achievement of students attending urban schools differed

significantly in kindergarten from students attending rural schools, and whether these same students differed in the rate at which they progressed during the 5 years of the study. We could also examine other characteristics of the students. For example, we could examine the rate at which students with one specific disability (e.g., autism spectrum disorder) progressed as compared to students with a different disability (e.g., learning disabled), and whether this observed relationship held for students in both urban and rural schools. Growth models are discussed in Chapter 3 of this manuscript, while cross-sectional models (i.e., one point in time) are discussed in Chapter 2.

The purpose of this paper is to introduce readers to the core concepts of HLM as applied to cross-sectional and longitudinal data. HLM is a complex topic and no assumptions are made about readers' familiarity with the topic outside of a basic understanding of regression. Thus, the bulk of this paper is dedicated to interpreting HLM analyses and important decisions that analysts make when building complex models. In Chapter 2, I begin with a brief explanation of nested data structures and some of the problems they pose. The primary components of a two-level model with cross-sectional data are then introduced and important elements of consideration discussed. In this section, the basic notation used for the null or unconditional model (no predictor variables) is introduced, as well as how it can be extended to include predictor variables. Model building and important statistics accompanying HLM analyses are also discussed, including overall model fit, the intraclass correlation coefficient (ICC), and the *Pseudo* R^2 statistic. All the basic concepts of HLM are introduced in this section, which is concluded with an illustrated example using real data.

The bulk of the paper is dedicated to Chapter 3, where the principles introduced for cross-sectional data are extended to illustrate how the concept of nesting can be used to measure

growth by treating *time* as nested within a *student*. In other words, just as many students may be nested within a school in a model with cross-sectional data, so too can multiple *test scores* be nested within an *individual* with longitudinal data. A two-level growth model is first introduced. It is then shown how the notation and model-building strategies can be expanded to three-levels. Considerable time is taken in Chapter 3 to reflect on specific issues related to growth modeling, including the linearity of the slope, the coding of *time*, and covariates that may vary by *time*.

It is important to note that this paper is intended to be educative for *consumers of research* – not for researchers intending to apply the techniques. However, HLM notation is discussed in considerable detail given that it is critical to understanding the specific model that has been applied. As stated earlier, it is also assumed that readers have a basic understanding of regression. For readers not yet familiar with regression, I recommend reading the first two sections of the structural equation modeling manuscript from this series (Anderson, Patarapichayatham, & Nese, 2013). Because the intended audience of this paper is consumers of research, and not researchers, there will be some issues that will be covered in less depth than interested readers may prefer. For those interested in more detailed accounts of the topics presented here, I recommend Raudenbush and Bryk (2002), Hox (2010), and (Snijders & Bosker, 2011) for a comprehensive overview of HLM and Singer and Willett (2003) for more detail surrounding analyses with longitudinal data. Finally, it is worth mentioning that readers completely unfamiliar with HLM may not follow every detail of the illustrated examples. One should not be concerned if this is the case. I opted to be overly inclusive in the examples, rather than gloss over important elements of the analysis, reasoning that more advanced readers may be interested in the model building decisions.

Chapter 2: Basic HLM Concepts

In contexts where data are nested (e.g., students nested within schools) and the effect of predictor variables on some outcome depend on that nesting, it is important that the nesting be accounted for in the model to avoid misrepresenting the effects. In many cases, individuals nested within some higher-level unit are more similar to each other than they are different and the observed effect of some treatment may then depend, in part, upon their membership to a specific higher-level unit (e.g., students attending School A instead of School B). If the data are dependent upon the higher-level unit then the residuals of individuals within the unit will be correlated – violating the independence of observations assumption of standard single-level regression (Raudenbush & Bryk, 2002).

As Roberts (2004) showed, not accounting for nested structures may potentially have dramatic effects and can even reverse the fundamental findings of the study. Roberts used a composite variable called "urbanicity" to predict students' science achievement. Without accounting for the nesting there was a .77 correlation, indicating that as students' urbanicity score increased, their science achievement generally did as well. This finding was puzzling and counter to previous research (e.g., Hannaway & Talbert, 1993). However, by accounting for the nesting of students within schools, the correlation became -.81, implying that as students' urbanicity increased their science achievement generally decreased. The reversal of the observed effect was found because *within* each school the correlation was negative, but by ignoring the context of the school the correlation "looked" positive. Generally, accounting for nesting does not produce as dramatic results as those shown by Roberts, but it serves as a good reminder for why accounting for nesting is so important.

Not accounting for nested data structures generally leads to aggregation bias, misestimated standard errors, and heterogeneity of regression (Raudenbush & Bryk, 2002). Aggregation bias occurs when a variable takes on a different meaning in its aggregated form than it does in its disaggregated form. When entering the aggregated variable into the model as a predictor variable it may then bias the results toward the aggregated meaning of the variable. As Raudenbush and Bryk highlight, for example, the overall composition of a school's student population (e.g., percent of students eligible for free or reduced price lunch; FRL) may have an effect on individual students beyond the effect of *individual* FRL eligibility. As will be discussed later in the paper, HLM can readily handle these complex relationships.

Misestimated standard errors occur when we fail to account for the dependence upon the higher-level units – i.e., when the independence of observations assumption is violated. HLM corrects the estimation by including the higher-level units in the model so that observations within a unit are independent. Heterogeneity of regressions occurs when the relation between a predictor variable and a specific outcome vary by some higher-level unit. For example, perhaps in some schools the relation between FRL and student achievement is quite strong while in others it is considerably weaker. Standard single-level regression would ignore this heterogeneity and assume the relation is constant across schools, while HLM can explicitly test and account for the heterogeneous relationships.

Two-Level Models

In this section I begin by providing an overview of the notation used in HLM for a two-level model. I then move to a section on model building. Model building in HLM must be systematic and theoretically based. The complexity of the models necessitate careful

consideration of each decision made so the final model makes sound theoretical sense and the analyst does not "overfit" the model to the specific sample he or she has attained.

Notation

The notation for all HLM models can be displayed in two ways: by the level of analysis, or in a single equation called a "mixed model". Models that are not terribly complex are often displayed as a single equation. However, as the models become more complicated it is often helpful to have the equations separated by the level of analysis. Equation 2.1 below details a basic two-level HLM model with no predictor variables, displayed by level.

$$\begin{cases} Y_{ij} = \beta_{0j} + r_{ij} \\ \beta_{0j} = \gamma_{00} + u_{0j} \end{cases} \quad (2.1)$$

At level 1, Y_{ij} represents the outcome *Y* for level one unit *i* nested in level two unit *j*, and is equal to a level one intercept, β_{0j}, and residual or unexplained variance r_{ij}. At level 2, the level 1 intercept, β_{0j}, is set as the outcome in a new regression equation with two components: the level 2 intercept, γ_{00}, and a random parameter, u_{0j}, which is the level 2 residual variance. The level 2 random parameter, u_{0j}, is what allows the model to vary by the higher-level unit. Equation 2.1 can also be displayed in its mixed model form, by simply substituting the level 2 equation into the level 1 equation. We then obtain:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \quad (2.2)$$

Note that Equation 2.2 represents the same model as Equation 2.1, but as a single equation. The β_{0j} term has now dropped out of the model given that it is defined by $\gamma_{00} + u_{0j}$. In other words, Equation 2 simply represents the substitution of β_{0j} with the higher-level terms that define it.

To better illustrate what these terms mean we can utilize a typical example in education – students nested within schools. We can use HLM to control for this nesting and more accurately

examine the effects of specific predictor variables on an outcome. For instance, say we were interested in the math achievement of students. Equation 2.2 then becomes

$$Math_{ij} = \gamma_{00} + u_{0j} + r_{ij} \tag{2.3}$$

Note that all that has changed at this point is that the outcome Y_{ij} has been substituted with $Math_{ij}$. However, now that we have some context of a problem we can better define what the terms mean. In this model, the math achievement of student *i* nested in school *j* is equal to the average achievement of schools (i.e., school level intercept), γ_{00}, plus the random component related to the school the student attends, u_{0j} (i.e., the difference between the overall average school achievement and average achievement for school *j*, the school the student attends), plus residual variance unique to the student and not captured by the model, r_{ij}. The r_{ij} term includes all the "left over" variance in the model, and includes measurement error, unaccounted for variables in the model, or potentially a host of other factors. The random component u_{0j} is what differentiates HLM from standard single-level regression because it allows the intercepts of schools to vary, whereas in single-level regression only one intercept would be calculated and assumed fixed, or equal, across schools. HLM relaxes this assumption, and allows school intercepts to vary randomly (in other words, the intercepts are freely estimated).

Predictor variables can also be added to the model at level 1, level 2, or both. Adding one predictor to each level results in the following model

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \end{cases} \tag{2.4}$$

Where X_{ij} represents a predictor variable for individual *i* nested in level 2 unit *j*, and W_j represents a predictor variable for level 2 unit *j*. Note that for each new predictor added to the model at level 1 we get a new beta at level 2 that is set as an outcome. In other words, each term

in the level 1 model has its own regression equation at level 2, which can include both an intercept and a residual (random effect). The random effects at the higher levels in the model (u_{ij}) can also be fixed at 0, which forces the effect to stay constant across all level 2 units, as with single-level regression. Indeed, if all random effects at the higher levels were fixed to 0 then the equation simplifies to a single-level regression equation. Thus, the entire basis of HLM lies in the random effects at the higher levels. For each of these variables researchers have to determine whether it makes theoretical sense to include the random effect for the term or not. If there is no reason to believe that any of the effects will vary at the higher levels – including the intercept and all predictor variables – then HLM may not be warranted. However, as we will see later we can also use the null model in Equation 2.1 to test for the degree of dependence upon the higher levels by estimating the intraclass correlation coefficient.

Now that the model has become more complex it is also critical that we pay close attention to subscripts so we can recognize what each term is referencing. In a two-level model each term has two subscripts, the first of which corresponds to level 1 while the second refers to level 2. For each subscript, 0 refers to an intercept, while all other numbers simply represent a sequential count of predictor variables added to the model. This can quickly become confusing, as γ_{00}, γ_{01}, γ_{10}, and γ_{11} all refer to different parameters. With practice, however, it can eventually become second nature. For example, γ_{00} is the overall intercept in the model, as there are no predictor variables included. The γ_{10} term represents the coefficient for the first predictor at level 1, while γ_{01} refers to the first level 2 predictor variable *of* the level 1 intercept (i.e., 0 predictors at level 1). The γ_{11} is slightly more complicated, as it is the first level 2 predictor *of the first level 1 predictor*. For example, γ_{11} may refer to the effect of a student with a disability

(level 1 predictor) attending a private school (level 2 predictor) on the student's achievement (outcome).

We can, of course, also display Equation 2.4 in its mixed model form by substituting the terms from level 2 into level 1, as follows:

1) Equation 2.4 is defined as

$$\begin{cases} Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \end{cases}$$

2) Substitute in $\gamma_{00} + \gamma_{01}W_j + u_{0j}$ for β_{0j}

$$\begin{cases} Y_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \beta_{1j}X_{ij} + r_{ij} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \end{cases}$$

3) Substitute in $\gamma_{10} + \gamma_{11}W_j + u_{1j}$ for β_{1j}

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + (\gamma_{10} + \gamma_{11}W_j + u_{1j})X_{ij} + r_{ij}$$

4) Redistribute

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}X_{ij} + \gamma_{11}W_jX_{ij} + u_{1j}X_{ij} + r_{ij}$$

By convention, the terms are generally rearranged so that the fixed effects appear first, followed by the random effects, which leads us to our final mixed model, defined as

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}W_jX_{ij} + u_{0j} + u_{1j}X_{ij} + r_{ij} \quad (2.5)$$

As predictor variables are added to the model it can quickly become quite complex. Paying close attention to the subscripts can help make the structure of the model clear so you can better understand whether the model makes theoretical sense, and what relationships specifically the researcher is testing.

Going back to our example with math, we may theorize specific predictor variables that affect student achievement at both the individual student level and the school level. For example,

we may hypothesize that students' eligible for free or reduced price lunch (FRL) may have significantly different math achievement than students not eligible for FRL. We may further hypothesize that FRL has a school level effect – as in, students in schools with a high proportion of FRL eligibility may have significantly different math achievement than students attending low proportion FRL eligible schools, independent of their own FRL status. Equation 2.4 then becomes

$$\begin{cases} Math_{ij} = \beta_{0j} + \beta_{1j}FRL_{ij} + r_{ij} \\ \beta_{0j} = \gamma_{00} + \gamma_{01}AverageFRL_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}AverageFRL_j + u_{1j} \end{cases} \quad (2.6)$$

There are a few important aspects to highlight in equation 2.6 corresponding to the underlying theory. First, the level 1 model is quite basic, simply stating that the math achievement of students depends, in part, upon their FRL status. The level 2 model, however, is a bit more complex. First, the model states that a student's intercept depends, in part, upon the average FRL – or the proportion of FRL eligible students – in the school the student attends. This effect is specified to vary randomly across schools. But perhaps even more complex, the model states that the effect of an individual's FRL status upon his or her math achievement depends, in part, upon the average FRL of the school the student attends (the γ_{11} term). This effect is similarly specified to vary randomly across schools.

Model building

When building HLM models the researcher generally begins with a null, or empty model. Predictors are then added to the model in a forward or backward elimination approach. The overall fit of the model must also be considered throughout the model building process. Below, each of these issues is discussed in turn.

Null model. For a basic two-level model, the null model is defined simply by Equation 2.1. The model is equivalent to a one-way analysis of variance (ANOVA) and is used within an HLM framework primarily to establish a baseline model from which subsequent models can be compared, and to capture the degree to which variance at level 1 depends upon group membership at level 2. Dependence at the higher levels can be assessed through the intraclass correlation coefficient (ICC), defined as

$$\rho = \frac{\tau_{00}}{(\sigma^2 + \tau_{00})} \tag{2.7}$$

where,

ρ = the ICC,

$\tau_{00} = u_{0j}$ = variance at level 2

$\sigma^2 = r_{ij}$ = variance at level 1

The ICC ranges from 0 to 1.0 and describes the proportion of the total variance that depends upon group membership. Although not discussed yet, it is worth mentioning here that the ICC for three level models is calculated similarly, but simply includes the variance term for the third level in the denominator. The ICC can then be calculated for level 2, with τ_{00} in the numerator, or for level 3 with τ_{000} in the numerator.

Some (e.g., Lee, 2000) suggest interpreting the ICC as a means for determining whether HLM is warranted, or whether standard single-level regression would suffice. For example, if there is a small amount of dependence on the higher-level groupings then the independence of observations assumption of single-level regression may not be violated, and thus may be an appropriate technique. Yet, as Roberts (2007) suggests, small ICCs may not warrant abandoning HLM given that additional dependence can arise after predictors have been entered into the

model. The ICC, therefore, should be an initial indicator of the warrants of HLM, but small values should not immediately rule out its use.

Entering predictor variables. Predictor variables can be entered into an HLM analysis through a forward, backward elimination, or simultaneous "block-entry" approach. The choice of how to include predictors into the model often depends upon the a priori assumption about the relations between predictor variable (i.e., how they interact) and the overall purpose of the analysis. For example, if a researcher were including variables in a model that theoretically interacted in a meaningful way, he or she would likely want to enter both variables into the model simultaneously (along with, potentially, their interaction term). However, if the researcher was interested in the unique contribution of each predictor, independent of the other predictors in the model, then he or she would likely want to enter each predictor sequentially, examining the model fit between each subsequent model. Further, a researcher may be interested in, for instance, examining the effect of a specific predictor after a host of demographic variables have been controlled for. In this case, the researcher would enter all the demographic variables into the model (first block), run the analysis, then enter the predictor variable of interest in the model and rerun the analysis, testing for differences in model fit between the two models.

Centering. Perhaps more important than sequential or simultaneous entering of predictors, however, is the choice of centering. Variables can generally be entered into the model in three ways: uncentered, group centered, or grand-mean centered. The choice of centering changes the interpretation of the intercept, and improper centering can result in model misspecification and untrustworthy results. Centering is necessary whenever a variable does not have a true zero point. Just as with standard regression, the intercept represents the value on the dependent variable when all predictor variables are 0. When a variable does not have a true zero

point, the variable must be transformed or centered so that it does. For instance, scores on the SAT math test range from 200-800. Prior to analysis, 200 points from each score could be subtracted so the variable would have a true zero point, and the scale would range from 0-600.

When variables are entered uncentered, or a transformation such as that described above is used prior to the analysis, the variable maintains (roughly) its original scale. That is, the intercept represents the average score on the dependent variable for students with a 0 (or a transformed 0, i.e., 200 on the example above) on the predictor variable. Alternatively, the SAT variable could be entered group or grand-mean centered. Group centering refers to subtracting the average score from the higher-level group (e.g., school) for all students within said group. Thus, variables can only be entered group centered at the lower levels of the model, and not at the highest level. Group-centering also changes the model as a whole. That is, group-centering is not simply a linear transformation of the variable. The model fit overall will change when group-mean centering is used (Hox, 2010). When variables are entered group centered the intercept then represent the average score for the group for which the student is a member.

Finally, at any level variables can be entered grand-mean centered by subtracting the overall grand-mean (mean for the variable across all units) from each score. Grand-mean centering is a simple linear transformation of students' scores, and the model-fit is not changed. When variables are entered grand-mean centered the intercept represents the average score on the dependent variable for all individuals in the dataset. Differences in variable centering methods are important given that they change the interpretation of the intercept, and, potentially, the model overall. But centering can also have some additional benefits, such as reducing issues of multicollinearity, which can ease estimation.

Once variables have been entered into the model, we can estimate a "pseudo R^2" statistic. Unfortunately, there is no direct measure of the variance accounted for by HLM models – hence the name "pseudo". Pseudo R^2 statistics provide an indication of the amount of variance accounted for by comparing the variance component in an unconditional model to the same variance component in a conditional model. Pseudo R^2 can be calculated for the overall residual in the model, r_{ij}, or for any random parameter in the model (e.g., intercept variance). Pseudo R^2 is calculated by applying the following formula:

$$Pseudo\ R^2 = \frac{(\sigma^2_{unconditional} - \sigma^2_{conditional})}{\sigma^2_{unconditional}} \quad (2.8)$$

Applying Equation 2.8 provides an estimate of the proportional reduction in unexplained variance in the random parameter, accounted for by the predictor variables in the model. When exploring how predictor variables account for the variance in specific parameters, one would simply substitute the σ^2 terms for τ's.

Model fit. The primary fit statistic used in HLM analyses is the *deviance* statistic. The deviance statistic is equal to -2 * the natural log of the likelihood ratio. Note that the actual computation of the deviance statistic is based on the maximum likelihood estimation procedure (and thus can only be computed when maximum likelihood estimation is used), which is beyond the scope of this paper. However, it is worth noting that the deviance statistic is an incremental fit indicator, by which its values are only meaningful relative to the values obtained from other models. Yet, to compare the deviance between models, they must also be nested – meaning that one model could be constructed from the other by simply including or eliminating predictor variables. So, for instance, the deviance from a two-level model cannot be compared to the deviance from a three-level model, or a two-level model with a different outcome.

Part of the reason that we always begin model building with an unconditional, or null model – is so that we have a baseline from which to compare the deviance statistic to for subsequent nested models. In general, the researcher begins by examining the deviance statistic from the null model. Predictors are then entered at level 1, and the deviance for these conditional models are compared relative to the null model. Once a level 1 model has be "settled" upon, the researcher then proceeds to enter predictors at level 2, using the deviance from the final level 1 model for subsequent model fit comparisons. If the model includes more than two levels then the researcher continues this process until a final model is established (i.e., comparing the deviance from models with predictors at level 3 with the final level 2 model).

Deviance represents "lack of fit", with larger values indicating a poorer fitting model. The fit between two models can be statistically tested. The procedure is quite simple, given that the difference between two deviance statistics follows a chi-square distribution, with the degrees of freedom equal to the difference in the number of parameters estimated in the two models. If the resulting value is significant, then the model with the lower deviance value fits the data significantly better. Evaluating model fit can, at times, be a bit confusing given that individual coefficients added to the model may be significant, but the overall difference in model fit may not be significant. Theory, of course, should always guide decisions about subsequent steps in model building, but generally if the overall model does not fit significantly better, then parsimony would dictate the removal of the predictor variable in the more complex model, despite its significance.

Finally, I would be remiss if I failed to mention some of the underlying assumptions of HLM, which should be thoroughly evaluated before any results are interpreted. There are, in

general, six assumptions related to HLM – three concerning the error structure, and three concerning the predictor variables. These assumptions are outlined in Table 2.1 below.

Table 2.1

HLM Model Assumptions

Error structure assumptions	Predictor variable assumptions
Independent and normally distributed level 1 residuals, with a mean of 0 and common variance, σ^2.	Level 1 predictors independent of level 1 residuals
Independent random effects at higher levels (i.e., level 2 & level 3), multivariate normally distributed, with a mean of 0 and a common variance, τ^2.	Higher level predictors independent of the residuals at the corresponding level
Residuals between levels are independent (i.e., no covariance between residuals at different levels).	Predictors at each level are independent of the random effects at other levels

While the researchers may not explicitly discuss model assumptions within a research paper, they should, at minimum, provide the reader with assurances that assumptions were investigated and that no major violations were found (perhaps through a footnote).

Illustrated Example 1: Two-Level Model

We will now move from discussing HLM from an abstract, theoretical position, to illustrating the technique through a series of concrete examples. We will begin by analyzing cross-sectional data in two ways - first with a single-level multiple regression analysis, then with a two-level HLM analysis. The analyses are conducted with the same dataset, and include students' specific disability type as a predictor of their state alternate assessment score for mathematics. The analyses are identical, with the exception of the HLM model accounting for the nesting of students within schools. Tables 2.3 and 2.4, below, illustrate the differences between the two models. Data for these analyses came from a sample of 288 students who had a

severe cognitive disability, were enrolled in fifth grade, and took the math portion of one state's alternate assessment during the 2010-2011 school year. All students were identified as having one of seven specific disabilities: (a) intellectual disability, (b) communication disorder, (c) emotionally disturbed, (d) orthopedic impairment, (e) other health impairment, (f) autism spectrum disorder, or (g) learning disabled. The referent group for both analyses was students with a learning disability. Table 2.2 displays descriptive statistics for each variable.

Table 2.2

Descriptive Statistics: Example 1

Variable	N	Mean	Standard deviation
Intellectual disability (ID)	77	.27	.44
Communication disorder (ComDis)	36	.13	.33
Emotionally disturbed (ED)	8	.03	.16
Orthopedic impairment (OI)	13	.05	.21
Other health impairment (OHI)	35	.13	.33
Autism spectrum disorder (Aut)	62	.22	.41
Learning disabled (LD)*	57	.20	.40
Students' scores on the state alternate assessment (MthRIT)	288	97.87	15.17

*Referent group

For both analyses, MthRIT represented students' scores on the math portion of the alternate assessment. All disability variables were dummy-coded vectors, coded 1 if the student had the disability, and 0 otherwise. Table 2.3 displays the results from the single-level multiple regression analysis.

Table 2.3

Single-Level Regression Results: Example 1

	Unstandardized coefficients		Standardized coefficients			Correlations		95% Confidence interval	
Variable	*b*	*SE*	*β*	*t*	*p*	Zero-order	Semi-partial	Lower bound	Upper bound
Intercept	108.75	1.73		62.93*	.00			105.35	112.15
ID	-15.93	2.28	-.47	-6.99*	.00	-.20	-.39	-20.42	-11.44
ComDis	-3.37	2.78	-.07	-1.21	.23	.19	-.07	-8.84	2.10
ED	-1.82	4.93	-.02	-.37	.71	.10	-.02	-11.52	7.88
OI	-27.49	4.01	-.38	-6.85*	.00	-.24	-.38	-35.38	-19.59
OHI	-8.64	2.80	-.19	-3.08*	.00	.06	-.18	-14.16	-3.13
Aut	-17.89	2.39	-.49	-7.47*	.00	-.24	-.41	-22.60	-13.18

* $p < .05$

The results of the model suggest that approximately 21 percent of the variance in alternate assessment scores was accounted for by students' disability category ($R^2 = .21$). Because students with LD were entered as the referent group, the intercept of the regression equation represents the average test score for students with LD on the state alternate assessment for math (108.75), as shown in the column labeled *b* of Table 2.3. Below the intercept, the effect of each predictor variable (student disability type) is reported. For example, students with an intellectual disability scored, on average, 15.93 points lower than students with LD. Similarly, students with autism spectrum disorder scored, on average, 17.89 points lower than students with LD. All predictor variables were significant ($p < .05$), with the exception of students identified with an emotional disturbance or communication disabled.

When conducting the HLM analyses, we began by specifying the unconditional, or null model, by specifying *MthRIT* as the outcome variable in Equation 2.1 to get

$$\begin{cases} MthRIT_{ij} = \beta_{0j} + r_{ij} \\ \beta_{0j} = \gamma_{00} + u_{0j} \end{cases} \tag{2.9}$$

This model was, again, defined primarily to (a) evaluate the degree to which *MthRIT* scores depend upon the school the student was nested within, and (b) provide a baseline deviance statistic from which subsequent models could be compared. The results of the model displayed in Equation 2.9 are displayed in Table 2.4 as Model 1. The level 1 variance (σ^2) was 112.28, and the level 2 variance (τ_{00}) was 115.25. Plugging these numbers into Equation 2.7, we get

$$\rho = \frac{115.25}{115.25+112.28} = .5065 \tag{2.10}$$

which equals the proportion of variance that lies between schools. Thus approximately 50.65% of the variance in *MthRIT* scores depended upon the school students attended (note: this is an unusually high number). The model had 2 estimated parameters, with deviance = 2246.932195.

The next step in model building process was to add predictor variables at level 1. For this model, it would make little sense to enter the predictor variables separately given that they were a set of dummy-coded vectors. That is, the set of variables together represent a single theoretical construct (*disability*). Thus, all variables were entered into the model simultaneously, as

$$\begin{cases} MthRIT_{ij} = \beta_{0j} + \beta_{1j}(MR_{ij}) + \beta_{2j}(ComDis_{ij}) + \beta_{3j}(ED_{ij}) + \beta_{4j}(OI_{ij}) + \\ \beta_{5j}(OHI_{ij}) + \beta_{6j}(Aut_{ij}) + r_{ij} \\ \\ \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} \\ \beta_{2j} = \gamma_{20} \\ \beta_{3j} = \gamma_{30} \\ \beta_{4j} = \gamma_{40} \\ \beta_{5j} = \gamma_{50} \\ \beta_{6j} = \gamma_{60} \end{cases} \tag{2.11}$$

Note that the only difference between Equation 2.11 and the single-level model is the random effect u_{0j} on the intercept, allowing the term to vary by school. The results of the model shown in Equation 2.11 are displayed in Table 2.4 below as Model 2. Unfortunately, we cannot test for statistical significance in the difference between the deviance statistics between Model 1 and Model 2 because there were not sufficient degrees of freedom. However, as shown in Table 2.4, the deviance statistic does reduce noticeably. We can also explore the proportional reduction in unexplained variance by applying Equation 2.8, as follows

$$Pseudo\ R^2 = \frac{(112.28127 - 101.33984)}{112.28127} = 0.097 \tag{2.12}$$

Thus, the inclusion of students' disabilities reduced the unexplained variance by approximately 10%. Similarly, we can explore the variance in the intercept accounted for by the predictors by

$$Pseudo\ R^2 = \frac{(115.25220 - 76.50049)}{115.25220} = 0.336 \tag{2.13}$$

which leads us to determine that students disability accounts for approximately 33.6% of the variance in students' intercepts. Notice that all the values within Equations 2.12 and 2.13 can be found in Table 2.4. The values are simply carried out to more decimal places in the equations.

If, for some reason, we hypothesized that the effect of a specific disability type on *MthRIT* depended on the school the student attended, we could simply estimate the corresponding random effects, as displayed in Equation 2.14.

$$\begin{cases} MthRIT_{ij} = \beta_{0j} + \beta_{1j}(MR_{ij}) + \beta_{2j}(ComDis_{ij}) + \beta_{3j}(ED_{ij}) + \beta_{4j}(OI_{ij}) + \\ \beta_{5j}(OHI_{ij}) + \beta_{6j}(Aut_{ij}) + r_{ij} \\ \\ \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \\ \beta_{2j} = \gamma_{20} + u_{2j} \\ \beta_{3j} = \gamma_{30} + u_{3j} \\ \beta_{4j} = \gamma_{40} + u_{4j} \\ \beta_{5j} = \gamma_{50} + u_{5j} \\ \beta_{6j} = \gamma_{60} + u_{6j} \end{cases} \quad (2.14)$$

Again, because the set of dummy-coded variables represent a single theoretical entity, it would make most sense to estimate all the random effects or none of the random effects. The results of Equation 2.14 are displayed as Model 3 in Table 2.4.

Table 2.4

HLM Results: Example 1

Fixed Effects	Model 1	Model 2	Model 3
Intercept, γ_{00}	97.76**	105.55**	108.68**
ID, γ_{10}	-	-11.21**	-15.89**
ComDis, γ_{20}	-	-4.34**	-3.60**
ED, γ_{30}	-	-2.51*	-2.00
OI, γ_{40}	-	-18.24**	-15.43**
OHI, γ_{50}	-	-6.22**	-9.49**
AUT, γ_{60}	-	-11.60**	-14.32**
Variance Components (Random Effects)			
Within-student, r_{ij}	112.28	101.34	68.51
Intercept, u_{0j}	115.25**	76.50**	0.17
ID, u_{1j}	-	-	143.72
ComDis, u_{2j}	-	-	0.85
ED, u_{3j}	-	-	0.17
OI, u_{4j}	-	-	328.08
OHI, u_{5j}	-	-	99.35
AUT, u_{6j}	-	-	176.11
Deviance	2246.93	2181.25	2126.61

$^{**}p < .01$, $^{*}p < .05$

There are a couple of things to note about the HLM results. First, thus far only the results for models with predictors at level 1 (note the 0 in the second subscript in the fixed effects

notation) have been presented. Second, although Model 3 technically fits the data the best, as it has the lowest deviance statistic, it displays some odd results. For example, the variance component values are both very high and very low, depending on the variable. Further, none of these values are statistically significant – suggesting the effect of student disability does not depend on school. Finally, Model 3 is perhaps less theoretically plausible than Model 2. That is, one would expect the effect of a specific disability type to be largely invariant across schools, and not depend upon school membership, given that it is a trait of the student. While instruction occurring within the school may influence this relationship, that instruction is mostly accounted for by the random intercept term.

If we consider Model 2 our final model, and compare the HLM results in Table 2.4 to our single-level results in Table 2.3, we can see some important differences. The intercept is approximately 3 points lower, while the magnitude of the effects of intellectual disability, orthopedic impairment, other health impairment, and autism are all markedly reduced. The effects of all remaining disability variables, however, actually increase in magnitude. These differences arise purely by accounting for the nesting of students within schools.

Although we conclude the discussion of cross-sectional data here, it is also important to highlight that any number of school-level predictors could be added to the model at level 2. These variables could be entered as a predictor of the intercept or any of the level 1 predictors – analogous to a cross-level interaction (which is evident when examining the full mixed model equation). For example, we may hypothesize that students' *MthRIT* score depends upon the size of the school. A school size variable could then be entered as a predictor of the level 1 intercept as $\gamma_{01}(Size)$. If we theorized that, for some reason, the effect of a students' disability – say ED – on *MthRIT* depended on school size, we could easily model the interaction with $\gamma_{31}(Size)$. Note

how the subscripts change between these variables. In both cases *Size* is the first level 2 predictor (second subscript), but in the first example *Size* is a predictor of the intercept (i.e., 0 predictors in the model) and in the second example *Size* is a predictor of the third level 1 predictor, ED. Thus, the flexibility in the types of relations HLM allows researchers to model is tremendous.

Chapter 3: Growth Models

The concept of nesting – i.e., students nested within schools – can be readily applied to the study of *change*. In education we are, of course, often interested in how students are changing. For instance, we may want to know simply the rate that students are acquiring a given skill. Or, similarly, we may want to know if students exposed to a particular intervention learn at a differential rate than their peers not exposed to the intervention. Historically, measuring change has posed both a statistical and psychometric challenge (see Harris, 1962), to such an extent that some have even advised against it in its entirety (Cronbach & Furby, 1970). Yet, persistent questions of change lingered and researchers were not content to simply "frame their questions in other ways" (Cronbach & Furby, p. 80). The 1980's saw an influx of new statistical models more apt to handle the challenges of measuring change, including HLM (Singer & Willett, 2003).

The primary challenges to studying change over time are issues related to research design, measurement, and statistics. From a research design perspective, researchers often only obtain data at one or two occasions (Raudenbush & Bryk, 2002). One obviously cannot study change with just a single measurement occasion, but even two occasions leads the analyst hamstrung. Under certain conditions a gain score can be calculated reliably (Zimmerman & Williams, 1998), but the reliability of the slope increases quite dramatically as additional testing occasions are collected (Willett, 1989). It is also worth noting that analysts will, on occasion, attempt to draw inferences of growth from a single measurement occasion. For instance, one may collect a single data point from sixth graders, seventh graders, and eighth graders all at the same time, then try to draw inferences on how students progress over time. Yet each of these groups of students are distinct, and any differences in scores may have just as much to do with the students themselves as with any change that has occurred.

Measurement issues are largely beyond the scope of this paper, but it is worth noting that, unsurprisingly, the quality of the growth estimates depend heavily upon the quality of the measures used. From a statistical perspective, the primary challenge to measuring growth is correlated residuals. That is, if students take a test at one point in time, the residual variance from those scores is likely to be correlated with the residual variance from scores taken at a later time, making it difficult, if not impossible, to parse out "growth" from idiosyncratic characteristics of the students. However, if we view the testing occasions as nested within the individual, then we can control for which student the testing occasions are nested within and we are better able to control for these dependencies within the data (i.e., the correlated residuals) by calculating a different slope for each individual, rather than a single average slope across students, which we can use to evaluate changes over time. Again, however, one must first have multiple data points over time for each individual involved in the study.

One of the primary advantages of modeling growth through HLM is that it is quite flexible and assumes little about the data structure. Students could be administered a set of repeated measures with equal or unequal intervals between administrations, have missing data points at any occasion (or multiple occasions), and have irregular measurement schedules for each student within the study. Suppose, for example, that a researcher is interested in the growth that students make in math during one school year. Unfortunately for the researcher, all students in the sample were administered the measures in a seemingly erratic manner – some students received as few as 4 measures throughout the year while other received as many 12; some students were administered measures in the beginning of the year only, others at the end of the year only, and still others received a regular administration pattern for the duration of the school year. Despite this complex dataset, the analyst could estimate the average growth occurring

throughout the year without eliminating any students from the sample. The flexibility of data structures HLM can handle is of tremendous benefit, given that other techniques (e.g., single-level regression) would require discarding data to obtain a workable sample.

HLM is also flexible in the types of relations it allows the researcher to specify. Variables can be entered as predictors of students' intercept (i.e., where they begin) and/or their slope (i.e., their rate of growth), or they may vary by time. Further, because students are still nested within schools, we can easily expand the model to three levels. In a three level model, with *school* at level 3, school-level variables can be entered as a predictor of students' intercept or slope (e.g., public versus private), and/or students' growth trajectories can be used to estimate a "school effect" – that is, how students nested within a particular school differ in their growth relative to students nested in other schools. Thus, when repeated measures are available, HLM growth models provide a rich statistical framework from which substantively important questions can be explored.

In what follows, a basic overview of HLM growth models is provided. Changes in notation are outlined to distinguish HLM growth models from cross-sectional models, as suggested by Raudenbush and Bryk (2002). The functional form of the slope (i.e., linear, quadratic, cubic, etc.) is then discussed, along with how misspecification can threaten the validity of inferences from the analysis. The discussion then moves to predictor variables with growth models, including time-varying covariates, before concluding with two illustrated examples – one with two levels (*repeated measures* nested in *students*) and one with three levels (*repeated measures* nested in *students* nested in *schools*). As in Chapter 2, the same dataset is used for both examples. Given that three level models have yet to be discussed in this

manuscript, a section in Chapter 3 is also devoted to the changes in notation and interpretation for three level models.

Notation

Level one of a basic two-level model with a single predictor variable for cross-sectional data is given by

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + r_{ij} \quad (3.1)$$

as discussed in Chapter 1. For a growth model, the notation changes to

$$Y_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + e_{ti} \quad (3.2)$$

where the terms have essentially equivalent meanings. The subscripts change from *ij* to *ti* to represent *time* nested within *individuals*. The coefficients change from β's to π's, the predictor variables change from *X's* to *a*'s, and the residual term changes from an *r* to and *e*. The level 1 model is often referred to as the "within-person", "within-student", or "within-subject" model. The first a_{ti} variables are coded to represent *time* between measurement occasions.

To specify a linear slope we would include a single a_{ti} term, which simply represents the time between testing occasions. However, the term must also include a true zero point. For example, suppose we had annual state testing data on students from grades 3 to 8. We could specify a linear slope by

$$StateTest_{ti} = \pi_{0i} + \pi_{1i}(Grade - 3) + e_{ti} \quad (3.3)$$

where $(Grade - 3)$ represents the grade the student was enrolled in at the time the testing occasion occurred, minus 3 (so there is a true 0 point). In this case the intercept term, π_{0i}, would represent the average score of grade 3 students, while the first coefficient, π_{1i}, would represent the average yearly change in the outcome variable, $StateTest$. Within an HLM growth-modeling framework, the error structure (i.e., the e_{ti} values) is generally assumed normally

distributed with a mean of 0 and a common variance, σ^2. It is also important to recognize that estimation for a growth models is equivalent to estimation for cross-sectional data; the notation is simply changed to clearly distinguish the two.

Level two of an HLM growth model is often referred to as the "between-person", "between-student", or "between-subject" model, where each coefficient from level 1 is defined by its own regression equation. The level 2 model for cross-sectional data (i.e., the level 2 model for Equation 3.1) with predictor variables and random effects is defined as

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + \gamma_{11}W_j + u_{1j} \end{cases} \quad (3.4)$$

as discussed in Chapter 1. For a growth model, the notation changes to

$$\begin{cases} \pi_{0i} = \beta_{00} + \beta_{01}X_i + r_{0i} \\ \pi_{1i} = \beta_{10} + \beta_{11}X_i + r_{1i} \end{cases} \quad (3.5)$$

Notice that the level 2 growth model looks quite similar to the level 1 model for cross-sectional data. The similarities in notation are purposeful and helpful because in both types of models the β's represent coefficient at the person-level. As we will see when we move to a three-level growth model, γ's in growth models still represent coefficient at one level above the individual level. Thus, the π's simply represent the coefficients at a level *below* the person level. The referencing of subscripts, though changing slightly in notation, remains unchanged (i.e., the first subscript represents a sequential count of predictors at level 1, while the second represents a sequential count of predictors at level 2).

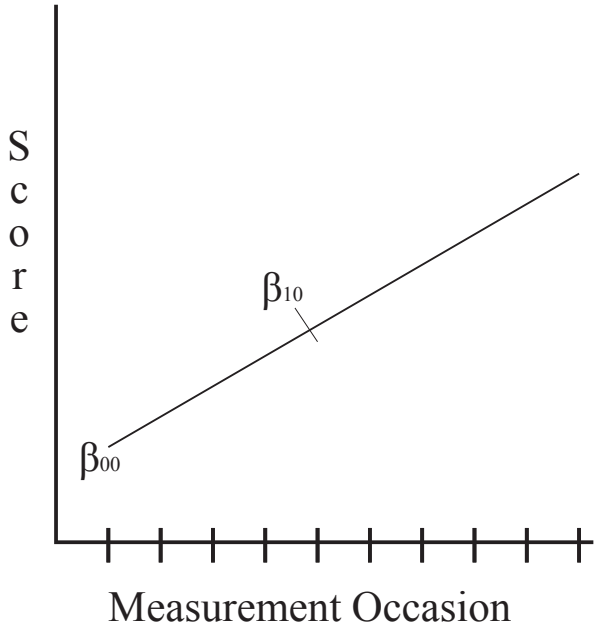
The specification of random effects as estimated or fixed is also important in growth models. Generally, researchers begin model building by specifying a "null growth" or "unconditional growth" model, with time entered as a predictor variable at level 1, and random effects estimated for both the intercept and the slope at level 2. That is,

$$\begin{cases} Y_{ti} = \pi_{0i} + \pi_{1i} a_{ti} + e_{ti} \\ \\ \pi_{0i} = \beta_{00} + r_{0i} \\ \pi_{1i} = \beta_{10} + r_{1i} \end{cases} \tag{3.6}$$

where all terms are defined as above. The conceptual underpinnings of the model displayed in Equation 3.6 will now be illustrated with a series of figures.

Suppose we were fitting a growth model for a small sample of 10 students. Figure 3.1 displays a hypothetical, single-level regression line for these students.

Figure 3.1 – Hypothetical Single-Level Growth Model

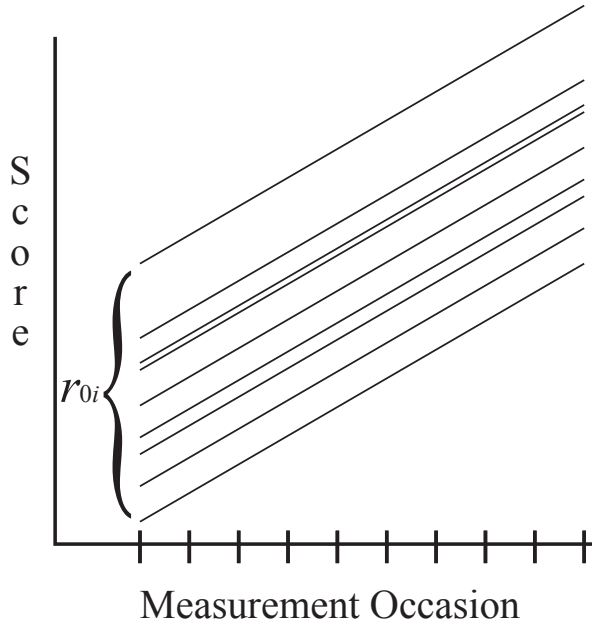


Notice that there is only a single line used to estimate the growth of all students. There is one intercept, β_{00}, which is fixed (i.e., estimated as equal) for all students and one slope, β_{10}, which is also fixed. For some students the line likely describes their growth pattern well, but for other students it is likely not an adequate representation. Figure 3.2 below displays a random intercepts model. That is

$$\begin{cases} Y_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + e_{ti} \\ \pi_{0i} = \beta_{00} + r_{0i} \\ \pi_{1i} = \beta_{10} \end{cases} \quad (3.7)$$

where each student's intercept is estimated based on his or her individual data.

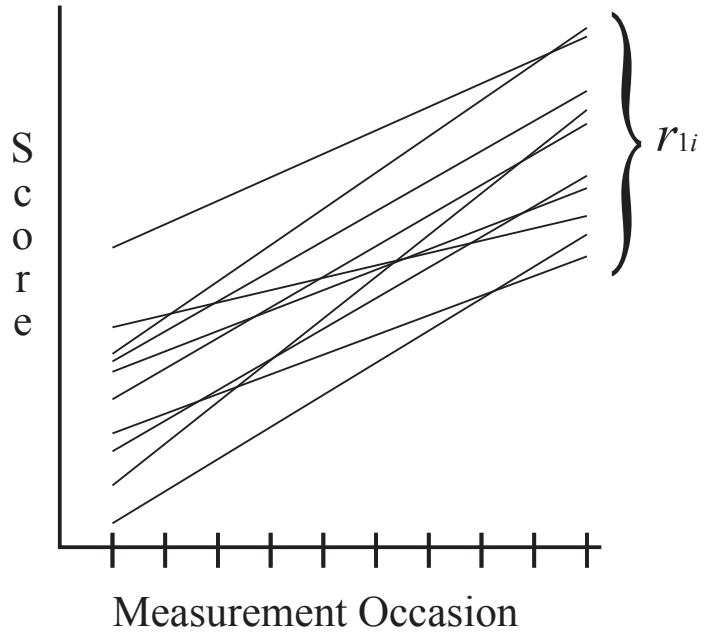
Figure 3.2 – Hypothetical Random Intercepts Model



Notice that the r_{0i} term is displayed in the figure to highlight that the random effect leads to each student's intercept being estimated individually. While the model displayed in Figure 3.2 likely estimates the observed data more accurately than the single-level regression line, it still assumes that the students progress over time at an equivalent rate.

A hypothetical random intercepts and slopes model – analogous to the null or unconditional growth model in Equation 3.6 – is displayed in Figure 3.3.

Figure 3.3 – Hypothetical Random Intercepts Model



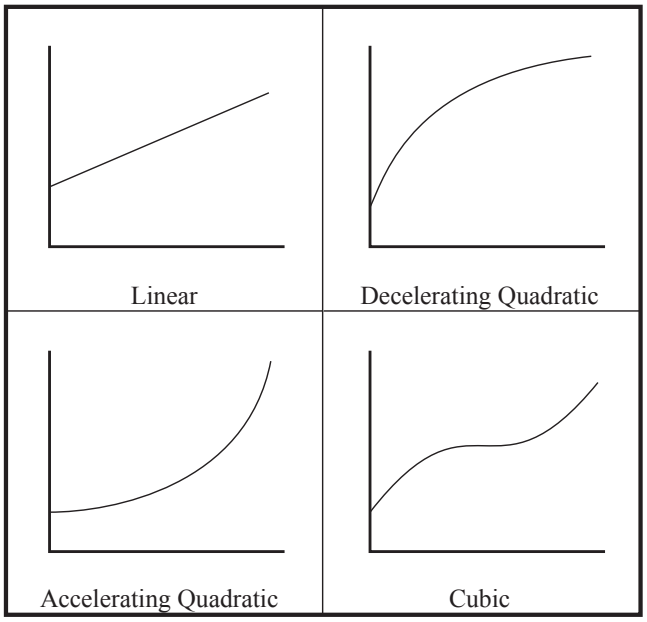
The random intercepts and slopes model allows each student's intercept and slope to be estimated based on his or her observed data. The r_{1i} term is highlighted in Figure 3.3 to highlight that the random effect estimates each students' slope. In most cases, the unconditional growth model in Equation 3.6, and displayed graphically in Figure 3.3, is the best initial estimate of students' growth given that it recognizes that students generally (a) begin with different achievement levels and (b) progress at differential rates over time.

Coding of Time and Functional Form

One of the primary threats to the validity of growth model inferences is the functional form the data follow. That is, do students progress at a constant rate? Or, do the data follow some other pattern, such as beginning slowly then rapidly increasing? Including higher-order polynomial time variables can test the functional form of the data. Figure 3.4 displays four commonly encountered functional forms: linear, decelerating quadratic, accelerating quadratic, and cubic. In educational data, linear and decelerating quadratic functional forms are the most commonly encountered (see, for example, Nese, Biancarosa, Anderson, Lai, & Tindal, 2011).

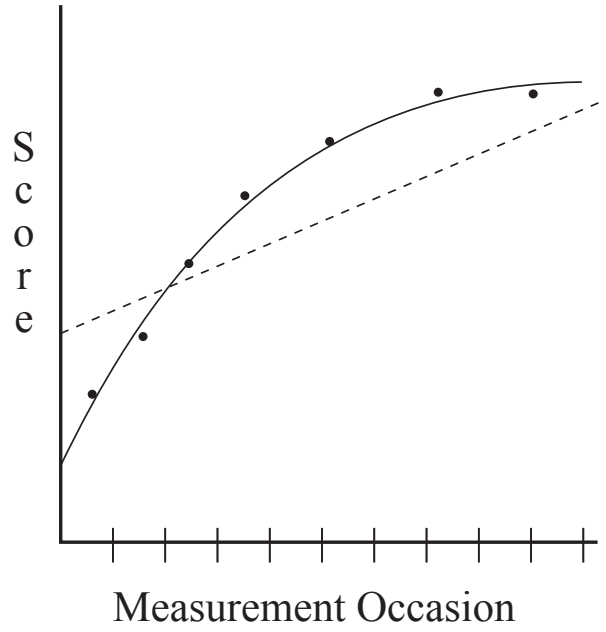
However, it is worth noting that all too often, researchers simply assume a linear functional form when higher-order polynomials may better model the data (for a discussion of functional form, see Singer & Willett, 2003, Chpt. 6)

Figure 3.4 – Functional Form



In nearly all cases, higher order functional forms should be tested to avoid making inappropriate inferences about the data. For example, Figure 3.5 below displays data that are clearly decelerating quadratic. If the data were fit by only a linear term (hatched line) the resulting inferences would be substantially different than if the data were modeled with a quadratic term (solid line). In this case, the linear term misrepresents the observed relation and the validity of any inferences made would be considerably threatened.

Figure 3.5 – Misspecified functional form



Yet, functional form can only be tested when there are sufficient data. When only three data points are available, a linear function generally must be assumed. When four data points are available, a quadratic term can be tested, and when 5 are available a cubic functional form can be tested.

In many cases, it may make the most sense to fit the data to a discontinuous functional form. That is, rather than fitting one continuous functional form to the entirety of the data, we could have a "break" in the middle to better represent the observed relation. If sufficient data points are available, functional form can be tested on each side of the discontinuity, with the functional form perhaps changing post-discontinuity. Depending on the way the data are coded, the discontinuity can be evaluated for changes in *level,* but not slope; changes in *slope,* but not level, or changes in *both* slope and level. Table 3.1 displays coding schemes for all the functional forms discussed thus far.

Table 3.1

Functional Form Coding Schemes

Student ID	Test Score	Grade	GrdLvl	Linear	Quadratic	Cubic	Post	RBa1	RBa2
1	100	3	0	0	0	0	0	0	0
1	103	4	0	1	1	1	0	1	0
1	104	5	0	2	4	8	0	2	0
1	108	6	1	3	9	27	1	2	1
1	108	7	1	4	16	64	2	2	2
1	109	8	1	5	25	125	3	2	3
2	97	3	0	0	0	0	0	0	0
2	102	4	0	1	1	1	0	1	0
2	104	5	0	2	4	8	0	2	0
2	103	6	1	3	9	27	1	2	1
3	109	5	0	2	4	8	0	2	0
3	112	6	1	3	9	27	1	2	1
3	113	7	1	4	16	64	2	2	2
…	…	…	…	…	…	…	…	…	…
i	y_{ti}	*t*		Grade - 3	Linear^2	Linear^3			

Note. All data are hypothetical. GrdLvl = Grade-level transition. Hatched line represent the point of discontinuity (grade-level transition). RBa1 and RBa2 represent a piecewise coding scheme suggested by Raudenbush & Bryk (2002) for fitting two separate growth functions pre- and post-discontinuity (with no baseline comparison).

Testing functional form. When testing for functional form, a "backwards elimination" approach is generally preferred. Backwards elimination includes the highest order polynomials of theoretical/empirical interest being entered into the model at the same time, with the highest order non-significant terms eliminated one by one, until a final functional form is settled upon. When testing for functional form, it is important to do so from *both* a theoretical and empirical basis. For example, in education we often observe a decelerating quadratic trend – the "learning curve". We thus may theorize that a quadratic term may need to be included. However, prior to fitting the model, it is important to investigate the observed data for each student, or a representative sample of students if the data set is large. If, by visual inspection, the data for many students appear to follow a quadratic trend, then it would be important to include the term in the model. However, if the majority of students appear to simply follow a linear trend, then parsimony may rule and the analyst would be justified in running a basic linear model.

As an example, imagine that we again are investigating the growth of students as the progress from grade 3 to 8 in the area of math. We have one test score per year. Assuming that theory and empirical evidence suggested that a quadratic model should be tested, the unconditional growth model would be defined as:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Grade - 3_{ti}) + \pi_{2i}(Grade - 3_{ti})^2 + e_{ti} \\ \quad \pi_{0i} = \beta_{00} + r_{0i} \\ \quad \pi_{1i} = \beta_{10} + r_{1i} \\ \quad \pi_{2i} = \beta_{20} + r_{2i} \end{cases} \tag{3.8}$$

Using backward elimination, we would first evaluate the quadratic fixed effect, π_{2i}. If the fixed effect were significant, then we would likely want to include it in the model. Next we would evaluate the quadratic random effect, r_{2i}. The random effect would indicate whether students differed significantly in their rate of deceleration (or acceleration). If the random effect were also significant then we would definitely want to keep the quadratic term in the model.

When both the random and fixed effects are significant, the model building decision-making process is quite simple. However, numerous complexities may arise. For instance, what if the quadratic terms are significant, but the linear terms are not? Should the linear terms be eliminated? The answer is no, because the linear term is needed to define the full quadratic curve. If the linear terms are not significant they still need to remain in the model, and the interpretation becomes that students do not have a significant *initial* rate of growth (i.e., it is not statistically different from 0), but that they decelerate (or accelerate) over time at a significant rate. It is also quite typical to observe a significant quadratic fixed effect, but not a significant random effect. In these cases, the random effect should likely be fixed before subsequent model building, with the interpretation being that there is an overall decelerating (or accelerating) trend, but the trend does not differ significantly between students. In cases where the quadratic random effect is significant but the fixed effect is not the decision as to whether to retain the parameter or not can be difficult, and should be made based on theory and other empirical sources of evidence (e.g., does the model overall fit the data significantly better than a linear model?).

When testing higher order function forms, such as cubic theory generally needs to play a more substantial role in educational data. For example, what phenomena would cause students to progress over time, slow down, and then progress again (as displayed in the lower right quadrant of Figure 3.4)? One likely occurrence would be the "summer-slide" (Cooper, Nye, Charlton, Lindsay, & Greathouse, 1996); however, drops due to summer are better coded with a discontinuous model to explicitly parcel out the time-period in which the students were not receiving instruction. It is worth noting, however, that the role that theory should play in guiding these decisions is a philosophical question, and reasonable and experienced analysts may

disagree on the best approach. An unconditional cubic growth model for our previous example would be defined as:

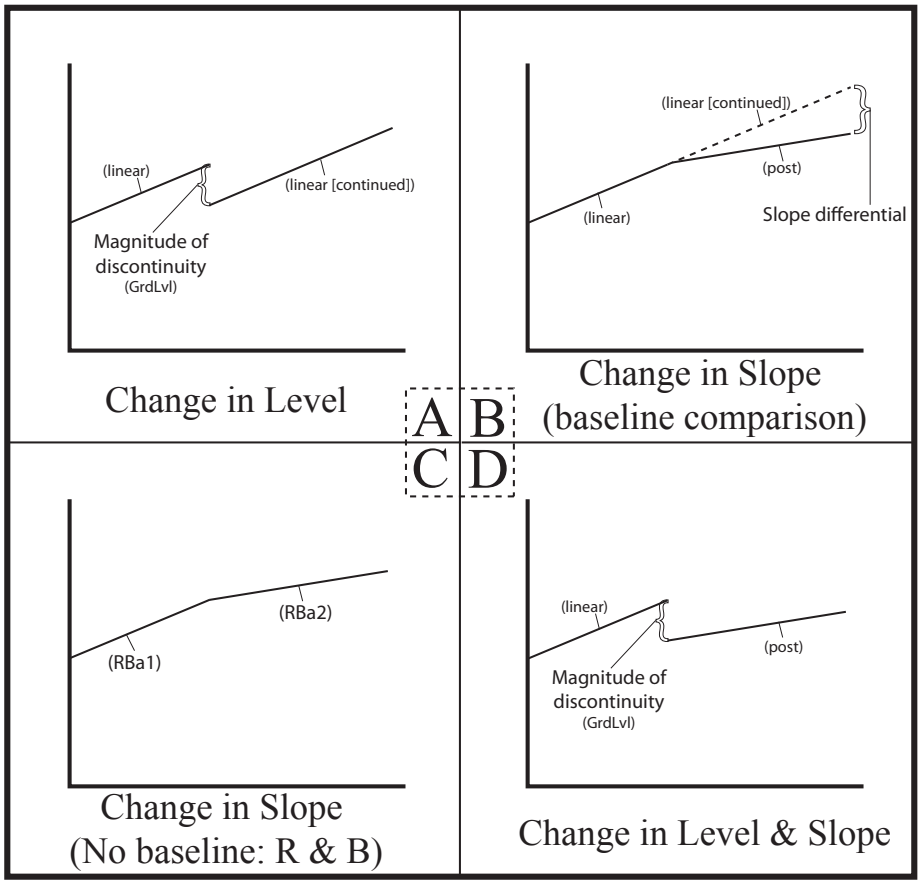
$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Grade - 3_{ti}) + \pi_{2i}(Grade - 3_{ti})^2 \\ \qquad\qquad + \pi_{2i}(Grade - 3_{ti})^3 + e_{ti} \\ \pi_{0i} = \beta_{00} + r_{0i} \\ \pi_{1i} = \beta_{10} + r_{1i} \\ \pi_{2i} = \beta_{20} + r_{2i} \end{cases} \tag{3.9}$$

The backward elimination model building process for a cubic model would be equivalent to the process described above for a quadratic model. Again, all terms would need to be included in the model, even if they were non-significant, up to the highest-order significant variable. If the cubic term was eliminated from the model, the process would start over again with the quadratic model. Backward elimination dictates that only one term is removed from the model at a time. So, if the model outlined in Equation 3.9 was run and both the quadratic and cubic terms were non-significant, the analyst employing backward elimination would *only* remove the cubic term before rerunning the model. The reason is because after eliminating the cubic term, it is quite possible for the quadratic term to become significant.

Discontinuous growth models. For discontinuous growth models, separate slopes are fit for different segments in the overall trajectory. There are three basic discontinuities that can occur: (a) changes in *level*, but not in slope; (b) changes in *slope*, but not in level; and (c) changes in *both* level and slope. Each of these are displayed in Figure 3.6. For a more in depth discussion on discontinuous models, see Singer and Willett (2003). There are many reasons one may theorize a discontinuity. For example, suppose in our example above that we theorized that students' entrance to middle school affected their growth. We may theorize that an initial drop in achievement would occur with the transition (i.e., change in level), or we may suspect that the overall rate of growth made in grades 6-8 would be slower relative to the growth made from

grades 3-5 (i.e., more complex skills lead to slower acquisition). Or, perhaps we may theorize that both a change in level and slope would occur with the transition to middle school. Each of these is a different research question with a different statistical model.

Figure 3.6 – Discontinuous Growth Models



Note. Texts in parentheses represent terms in the model, as coded in Table 3.1.

When evaluating a change in *level* at the point of discontinuity, one simply needs to include a time-varying covariate indicating the point at which the student entered middle school. The model would thus be defined as:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Linear_{ti}) + \pi_{2i}(GrdLvl_{ti}) + e_{ti} \\ \quad \pi_{0i} = \beta_{00} + r_{0i} \\ \quad \pi_{1i} = \beta_{10} + r_{1i} \\ \quad \pi_{2i} = \beta_{20} + r_{2i} \end{cases} \tag{3.10}$$

where *Linear* represents a basic linear growth slope coded across all grades, and *GrdLvl* is a dummy-coded vector representing the time the student entered middle school. If the *GrdLvl* variable was not significant, then no discontinuity in the data would be present and a linear trend would adequately model the observed relation across grades. However, if *GrdLvl* was significant, then an immediate change in level would occur coincidence with the middle school transition. In Figure 3.6A, a significant discontinuity is displayed. Note that the size of the discontinuity, as displayed in the figure, is equal to the magnitude of the π_{2i} coefficient.

When evaluating changes in *slope* at the point of discontinuity, but not level, one simply creates a second *time* variable coded by the number of years since transitioning to middle school, as displayed in Table 3.1 (Post). The discontinuous growth model would then be defined as:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Linear_{ti}) + \pi_{2i}(Post_{ti}) + e_{ti} \\ \quad \pi_{0i} = \beta_{00} + r_{0i} \\ \quad \pi_{1i} = \beta_{10} + r_{1i} \\ \quad \pi_{2i} = \beta_{20} + r_{2i} \end{cases} \tag{3.11}$$

The *Post* variable creates a slope for students' post-elementary. The slope of *Post* can then be compared to the baseline slope, *Linear*, to evaluate whether a significant difference in slope occurs in middle school. One possible version of Equation 3.11 is displayed in Figure 3.6B, with the *linear* slope post-elementary displayed with a hatched line, and the shallower *Post* slope displayed with a solid line. Note that, were the *Post* slope not significant, the slope differential (as shown in the figure) would not be significantly different than 0.

It is important to note that a common variant of equation 3.11 is often applied, as suggested by Raudenbush and Bryk (2002), where each segment of the slope is coded into two completely separate pieces. This coding scheme is displayed in Table 3.1 as RBa1, and RBa2. Under the Raudenbush and Bryk scheme, there is no baseline slope to compare to post-discontinuity, as displayed in Figure 3.6C. Thus, under Raudenbush and Bryk's suggested

coding, the π_{2i} term represents a completely new slope, while in Equation 3.11 it represents the difference between the linear slope and the post-discontinuity slope. The Raudenbush and Bryk scheme should, in most cases, be applied when there is a theory behind the mechanism that causes the discontinuity that creates an abrupt change in slope, and/or renders comparisons back to the slope pre-discontinuity largely not meaningful.

If, going back to our example, we theorized that the middle school transition would cause a change in *both* students' level and slope, the discontinuous growth model would be defined as:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Linear_{ti}) + \pi_{2i}(GrdLvl_{ti}) + \pi_{3i}(Secondary_{ti}) \\ \qquad\qquad + e_{ti} \\ \pi_{0i} = \beta_{00} + r_{0i} \\ \pi_{1i} = \beta_{10} + r_{1i} \\ \pi_{2i} = \beta_{20} + r_{2i} \\ \pi_{3i} = \beta_{30} + r_{3i} \end{cases} \tag{3.12}$$

where all terms are coded as in Table 3.1. If all terms were significant and our theory bore out, then the resulting fitted model may resemble the solid line in Figure 3.6D. For this model, our π_{0i} term would represent students' status at grade 3, the π_{1i} term would represent students' baseline growth from grades 3-8, the π_{2i} term would represent students' change in level at grade 6, and the π_{3i} term would represent students growth post-elementary. Using a backwards elimination approach, the model in Equation 3.12 would likely be the initial model fit to the data, provided the model made theoretical sense.

Three Level Models and Predictor Variables

To this point in the paper, only two level models have been discussed. For example, in Chapter 2 we discussed an example where students were nested in schools. In this chapter, we have discussed various two-level models where a set of repeated measures are nested within an individual. However, the basic concept of "levels" can easily be expanded to three, or even four

or more levels. For example, suppose we combined the two aforementioned examples. We would then have a three level model with repeated measures nested in students, who are nested in schools. Assuming the outcome of interest was still math, and we were fitting only a basic unconditional linear growth model, the model would be defined by:

$$\begin{cases} MathScore_{tij} = \pi_{0ij} + \pi_{1ij}(Linear_{tij}) + e_{tij} \\ \\ \quad \pi_{0ij} = \beta_{00j} + r_{0ij} \\ \quad \pi_{1ij} = \beta_{10j} + r_{1ij} \\ \\ \quad \beta_{00j} = \gamma_{000} + u_{00j} \\ \quad \beta_{10j} = \gamma_{100} + u_{10j} \end{cases} \tag{3.13}$$

where all level two parameters are defined by their own regression equations at level three. Breaking the equation down into its parts, we can see that students' intercepts, π_{0ij}, are allowed to vary randomly between students, r_{0ij}, and between schools u_{00j}. Were the intercept not allowed to vary randomly at the higher levels, there would be no random terms, and $\pi_{0ij} = \beta_{00j} = \gamma_{000}$.

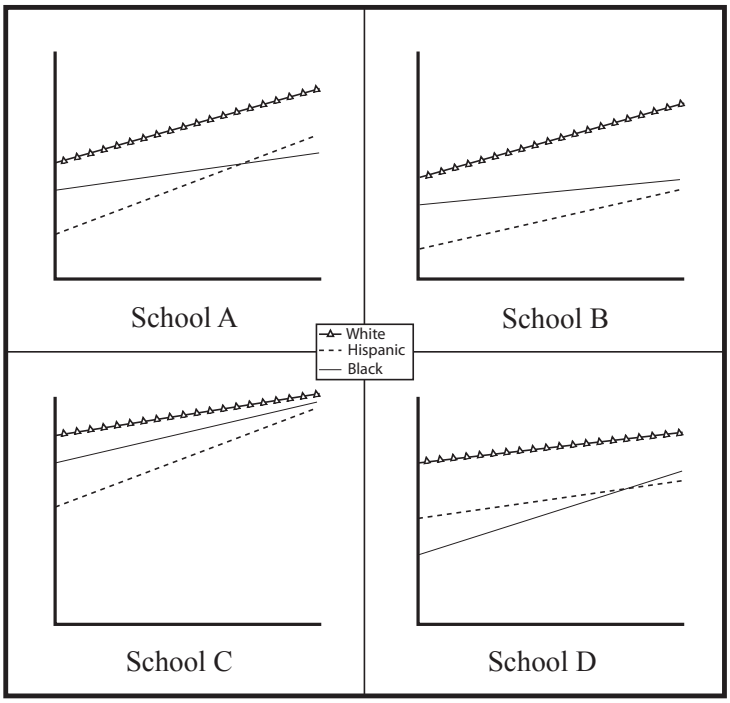
One of the most powerful features of modeling growth through an HLM framework is the flexibility provided when entering predictors. As previously mentioned, predictor variables can be added as a "time-varying covariate" at level 1 – implying that the effect of the specific variable does not occur until a specific point in time, or that the effect of the variable changes over time. Including time-varying covariates results in a discontinuous model where a change in level, but not slope, occurs. However, predictor variables can also easily be entered at the higher levels as well. For instance, suppose we wanted to explore the minority achievement gap (see, for example, Fryer & Levitt, 2004; Reardon & Galindo, 2009). We could enter dummy coded ethnicity variables called *Black* and *Hispanic* into the model as predictors of students' intercepts *and* slopes. We could then explore the differences in initial starting point (i.e., intercept) and rate

of growth over time (i.e., slope) for Black and Hispanic students relative to the reference group – White students. We could specify a three level model to explore whether the effect of these ethnicity variables on student achievement varied significantly between schools. The model would then be defined as:

$$
\begin{cases}
MathScore_{tij} = \pi_{0ij} + \pi_{1ij}(Linear_{tij}) + e_{tij} \\
\quad \pi_{0ij} = \beta_{00j} + \beta_{01j}(Black) + \beta_{02j}(Hispanic) + r_{0ij} \\
\quad \pi_{1ij} = \beta_{10j} + \beta_{11j}(Black) + \beta_{12j}(Hispanic) + r_{1ij} \\
\quad \beta_{00j} = \gamma_{000} + u_{00j} \\
\quad \beta_{01j} = \gamma_{010} + u_{01j} \\
\quad \beta_{02j} = \gamma_{020} + u_{02j} \\
\quad \beta_{10j} = \gamma_{100} + u_{10j} \\
\quad \beta_{11j} = \gamma_{110} + u_{11j} \\
\quad \beta_{12j} = \gamma_{120} + u_{12j}
\end{cases}
\tag{3.14}
$$

Notice that the model quickly becomes quite complicated, as we now have random effects for all predictor variables. The theory behind the model in Equation 3.14 is that Black and Hispanic students start out at different levels than White students, progress over time at a differential rate, and that these effects vary between schools. If all terms were significant, the resultant model may look something like Figure 3.7.

Figure 3.7 – Hypothetical Schools Displaying Differential Growth by Student Minority Groups



The figure displays the average growth rate of three student groups – White, Hispanic, and Black – nested within four schools (A through D). Notice that, for School A, Hispanic students (dashed line) begin at a lower achievement level relative to White students (triangles), but make more rapid growth over time. The White-Hispanic achievement gap is thus closing. However, Black students (thin solid line) begin with a lower level of achievement than White students, and progress at slower rate. In School A, then, the Black-White achievement gap is widening, at the same time the Hispanic-White achievement gap is narrowing. Yet, these relationships differ quite dramatically by school. For instance, in School C, both the Black-White *and* the Hispanic-White achievement gaps are narrowing. These complex relationships can all be modeled through Equation 3.14.

To illustrate the flexibility of HLM, we could also explore school-level predictor variables. That is, we could explore whether students attending a *private* school began at a different level, or progressed over time at a different rate, than students attending a *public* school

(reference group). We could also explore whether the effect of student ethnicity on student achievement depended on (i.e., was moderated by) the type of school the student attended. This model would be defined by:

$$
\begin{cases}
MathScore_{tij} = \pi_{0ij} + \pi_{1ij}(Linear_{tij}) + e_{tij} \\
\quad \pi_{0ij} = \beta_{00j} + \beta_{01j}(Black) + \beta_{02j}(Hispanic) + r_{0ij} \\
\quad \pi_{1ij} = \beta_{10j} + \beta_{11j}(Black) + \beta_{12j}(Hispanic) + r_{1ij} \\
\quad \beta_{00j} = \gamma_{000} + \gamma_{001}(Private) + u_{00j} \\
\quad \beta_{01j} = \gamma_{010} + \gamma_{011}(Private) + u_{01j} \\
\quad \beta_{02j} = \gamma_{020} + \gamma_{021}(Private) + u_{02j} \\
\quad \beta_{10j} = \gamma_{100} + \gamma_{101}(Private) + u_{10j} \\
\quad \beta_{11j} = \gamma_{110} + \gamma_{111}(Private) + u_{11j} \\
\quad \beta_{12j} = \gamma_{120} + \gamma_{121}(Private) + u_{12j}
\end{cases}
\tag{3.15}
$$

Note that the subscripts become increasingly important as the model complexity increases. For instance, γ_{121} refers to the first level three predictor, *Private*, of the second level two predictor, *Hispanic*, of the first level one predictor, *Linear*. This predictor is a cross-level interaction, and in English (rather than statistical variables), we would say that the term explores whether Hispanic students make significantly different growth in public versus private schools. Similarly, the γ_{011} term refers to the first level three predictor, *Private*, of the first level two predictor, *Black*, of the intercept. Again in English, this term tests whether Black students in private schools begin at a significantly different level than Black students in public schools.

Hierarchical linear models thus provide a rich basis from which complex phenomena can be modeled. In what follows, two illustrated examples are provided: one with a two-level growth model, and one with a three-level growth model.

Illustrated Example 1: Two-Level Growth Model

In the following example, the application of a two level HLM growth model is illustrated. The model was applied to evaluate the growth students made during grades 5 and 6 in the area of

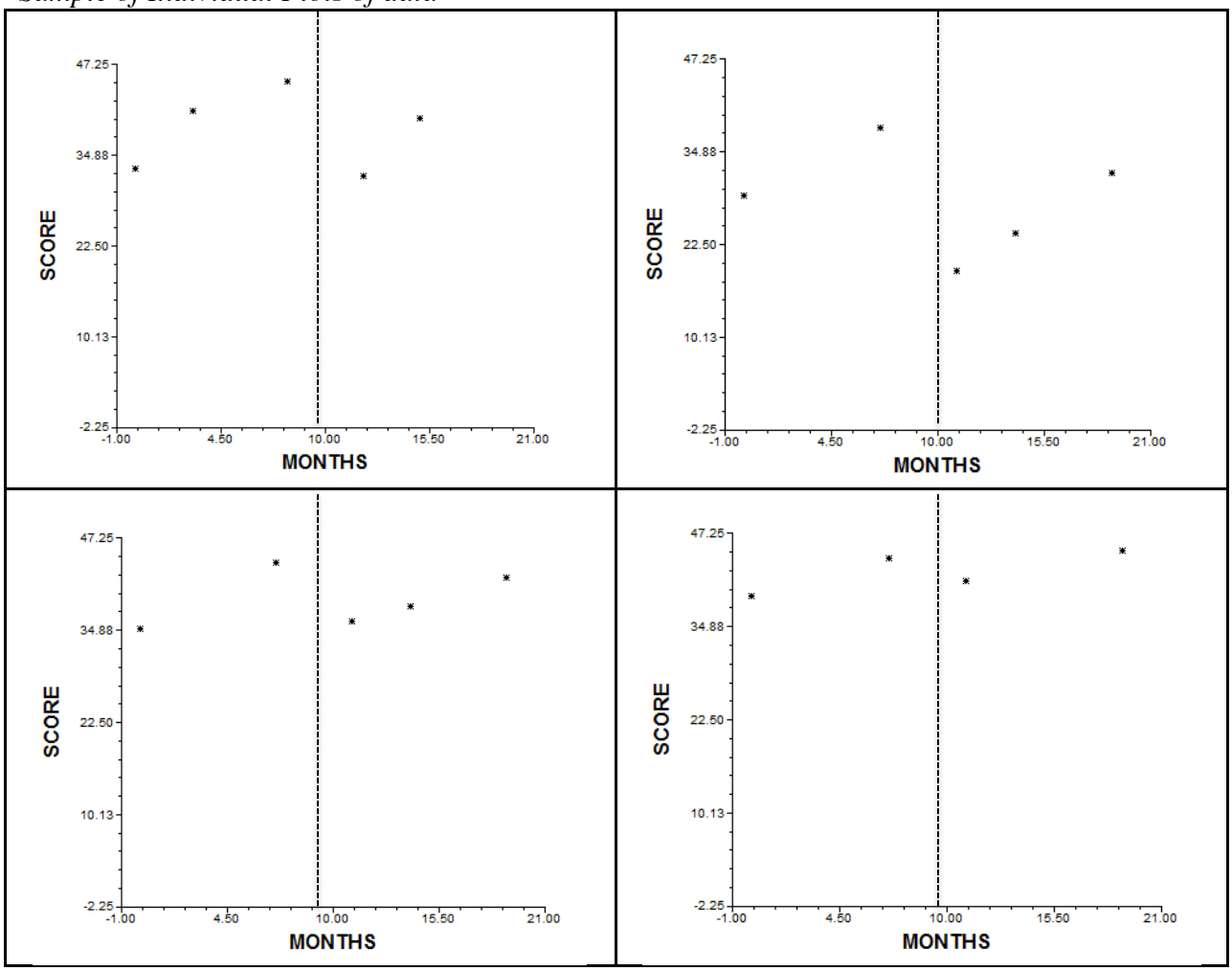
mathematics. There are a total of six data points – three from each year – which were obtained from an interim math screening assessment. The assessments used were all scaled to be of equivalent difficulty *within* each school year, but not across. Thus, to fit a growth model to all six time points, we have to apply a time-varying covariate that indicates the time at which students changed grades/test forms, analogous to the model displayed graphically in Figure 3.6A. However, in model testing we often set up a "straw man" from which comparisons can be made. We thus begin by testing a simple linear model to the entirety of the data, as follows:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Months_{ti}) + e_{ti} \\ \quad \pi_{0i} = \beta_{00} + r_{0i} \\ \quad \pi_{1i} = \beta_{10} + r_{1i} \end{cases} \tag{3.16}$$

For simplicity, it is assumed that a linear slope reasonably modeled the observed data within each year. The variable *Months* measures the time, in months, between testing occasions. The intercept represented students' score during the fall testing occasion in grade 5. Note that, because six time points were available, a model that followed an overall quadratic (or cubic) trend, but that had a discontinuity in level with the transition between grades, could have been tested. However, given that functional form was handled, in part, through the discontinuous design, and plots of the data for individual students suggested a linear trend reasonably modeled the data within each school year (as displayed in Figure 3.8 below) the more parsimonious model was chosen.

Figure 3.8

Sample of Individual Plots of data



Note. Figure displays sample plots of individual student data. Vertical line depicts the point of discontinuity.

The results of our sequence of models are displayed in Table 3.2. The model displayed in Equation 3.16 (Model 1 in Table 3.2) suggested that students began, on average, scoring 33.29 questions correct out of 45, and progressed at a rate of .07 questions correct per month. The model also suggests that students differed significantly in their intercepts, with a standard deviation of 5.47, but not in their slopes. However, the model is badly misspecified because we are not accounting for the change in test form difficulty that occurs with the change in grade level. The change in test form difficulty was modeled by including a covariate for grade-level in

the next model (coded 0, 0, 0, 1, 1, 1, for time points 1-6 respectively). The overall fit of the model was then compared to the "straw man" model displayed in Equation 3.16 via a chi-square deviance test. Despite its non-significance in Equation 3.16, the random effect for students' slopes in the model was maintained, given its theoretical importance. The model was thus defined as:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Months_{ti}) + \pi_{2i}(Grdlevel_{ti}) + e_{ti} \\ \quad \pi_{0i} = \beta_{00} + r_{0i} \\ \quad \pi_{1i} = \beta_{10} + r_{1i} \\ \quad \pi_{2i} = \beta_{20} + r_{2i} \end{cases} \tag{3.17}$$

The *Grdlevel* variable was specified as random, indicating that the anticipated dip in the overall achievement depended on the individual student.

The results of the model displayed in Equation 3.17 (Model 2 in Table 3.2) suggested that students began, on average, with a score of 32.18 points and gained, on average, 0.76 questions correct per month. The transition between grade levels coincided with an average drop of 10.85 points. All random effects were significant. Further, the chi-square deviance test was significant ($\chi^2 = 4819.32$, $df = 4$, $p < .001$), suggesting that the model displayed in Equation 3.17 (unsurprisingly) fit the data better than the model displayed in Equation 3.16.

Student demographic variables were then added to the model as predictors of each level one parameter. Four demographic variables were included: (a) special education status, (b) sex, (c) English language learner [ELL] status, and (d) free/reduced lunch status [FRL]. All variables were dummy-coded vectors, entered into the model uncentered. A backwards elimination procedure was followed, by which all predictors were added to the model simultaneously and evaluated together. The model was thus defined as:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Months_{ti}) + \pi_{2i}(Grdlevel_{ti}) + e_{ti} \\ \quad \pi_{0i} = \beta_{00} + \beta_{01}(SPED) + \beta_{02}(Female) + \beta_{03}(ELL) + \beta_{04}(FRL) + r_{0i} \\ \quad \pi_{1i} = \beta_{10} + \beta_{11}(SPED) + \beta_{12}(Female) + \beta_{13}(ELL) + \beta_{14}(FRL) + r_{1i} \\ \quad \pi_{2i} = \beta_{20} + \beta_{21}(SPED) + \beta_{22}(Female) + \beta_{23}(ELL) + \beta_{24}(FRL) + r_{2i} \end{cases} \quad (3.18)$$

The addition of the student covariates again resulted in a significantly better fitting model, (χ^2 = 846.53, *df* = 12, $p < .001$). All variables were significant predictors of students intercept, as displayed in Model 3 of Table 3.2, with, on average, special education students' intercept being 5.26 points lower than non-special education students, females scoring 0.64 points lower than males, ELL students scoring 2.88 points lower than non-ELL students, and students eligible for FRL scoring 4.07 points lower than non-FRL eligible students. *Female* was the only significant predictor of students' slope, progressing 0.12 questions correct per month faster than males (indicating a closure of the gender gap). Both Female and FRL were significant predictors of the effect of *Grdlevel* on students' math achievement. The coefficients suggested that females dropped 1.94 more than males between grade-levels and FRL eligible students dropped 1.40 points more than non-FRL students. Adding the demographic variables to the model accounted for approximately 26% of the variance in students' intercepts, 22% of the variance in students' slopes, and 8% of the variance in the effect of *Grdlevel* on math achievement.

From here, we could stop model building and report our results, or we could continue to refine the model to find the best fitting/most parsimonious model. Following the latter, we could remove all non-significant fixed effects. The model would then be defined as:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Months_{ti}) + \pi_{2i}(Grdlevel_{ti}) + e_{ti} \\ \quad \pi_{0i} = \beta_{00} + \beta_{01}(SPED) + \beta_{02}(Female) + \beta_{03}(ELL) + \beta_{04}(FRL) + r_{0i} \\ \quad \pi_{1i} = \beta_{10} + \beta_{11}(Female) + \beta_{12}(FRL) + r_{1i} \\ \quad \pi_{2i} = \beta_{20} + \beta_{21}(Female) + r_{2i} \end{cases} \quad (3.19)$$

The model displayed in Equation 3.19 once again fit the data significantly better than the previous, more complex model displayed in Equation 3.18 ($\chi^2 = 360.60$, $df = 10$, $p < .001$). All terms were significant, as displayed in Model 4 of Table 3.2.

Table 3.2

Parameter Estimates: Example 1

Fixed Effects	Model 1		Model 2		Model 3		Model 4	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Intercept, β_{00}	33.29	0.12	32.18	0.13	35.41	0.20	35.38	0.19
SPED, β_{01}					-5.27	0.34	-5.49	0.33
Female, β_{02}					-0.64	0.23	-0.67	0.23
ELL, β_{03}					-2.88	0.69	-2.48	0.62
FRL, β_{04}					-4.07	0.23	-3.91	0.22
Slope, β_{10}	0.07	0.01	0.76	0.01	0.69	0.02	0.70	0.02
SPED, β_{11}					-0.06*	0.03		
Female, β_{12}					0.12	0.02	-2.04	0.02
ELL, β_{13}					0.06*	0.08		
FRL, β_{14}					0.04*	0.02		
Transition, β_{20}			-10.85	0.14	-9.39	0.25	-9.46	0.22
SPED, β_{11}					0.69*	0.40		
Female, β_{12}					-1.94	0.29	-2.04	0.28
ELL, β_{13}					0.04*	0.96		
FRL, β_{14}					-1.40	0.29	-0.88	0.15
Random effects	Estimate		Estimate		Estimate		Estimate	
Within-student, e_{ti}	25.29		14.45		14.45		14.45	
Student intercept, r_{0i}	29.88		35.53		26.22		26.25	
Student slope, r_{1i}	0.01*		0.07		0.06		0.06	
Transition, r_{2i}			7.68		6.02		6.16	
Model deviance (parameters)	84820.70 (6)		80001.38 (10)		79154.86 (22)		79166.411 (17)	

Note. Transition = grade-level transition with new, more difficult measures. Coeff. = Model Coefficient

*coefficient is *not* significant, $p > .05$; All other values significant.

Model 1 = Equation 3.16
Model 2 = Equation 3.17
Model 3 = Equation 3.18
Model 4 = Equation 3.19

Overall, our final model (Model 4, Equation 3.19) suggested that students began, on average, with a score of 35.38 points. Students who received special education, were female,

English language learners, or were eligible for free or reduced price lunch all began significantly lower. Students progressed, on average, at a rate of 0.70 questions correct per month, with Females progressing significantly faster, gaining an additional 0.13 questions correct per month. Thus, while a sizeable initial gender gap exists, the results suggest that the magnitude of the gap decreases over time. However, no other student group progressed at a significantly different rate. When transitioning grade levels, students' scores dropped, on average 9.46 points. Female students experienced a greater drop, losing an additional 2.04 points, and students eligible for free or reduced price lunch dropped 0.88 points more than the reference group. Thus, the two groups with the largest disparities in initial achievement – females and students eligible for free or reduced price lunch – also experienced the largest drop in achievement with the transition between grade-levels. The gender gap therefore appears to be closing *during* the school year, but widening *between* school years. In the next section, the model is expanded to explore the school's impact.

Illustrated Example 2: Three-Level Growth Model

When exploring the impact of the school a student attends on his or her growth, one must consider how to handle mobile student data. All the models discussed thus far assume a "pure" nesting structure. That is, lower-level units were assumed to be members of one, and only one, higher-level unit. Yet, when students move between schools they become members of multiple groups. There are three general methods of handling mobile student data: (a) deleting all mobile students from the dataset, (b) attributing the entirety of the students' data as representative of the first or last school the student attended, or (c) applying a cross-classified model (Grady & Beretvas, 2010). Cross-classified models provide the most flexible method for handling mobile student data, but their estimation is considerably more complex and largely beyond the scope of

this paper. Both method A and B introduce some level of bias to the overall estimates, but are necessary to "create" a pure nesting structure. The degree to which these methods impact the overall inferences drawn depends largely upon the rate of student mobility, and the correlation between mobile student data and specific student demographics.

The aim of this paper is primarily to provide an overview of HLM, and illustrate its use in practice. Method B above was thus applied to obtain a purely nested dataset, although in practice a cross-classified model would likely better represent the overall structure of the data. Extending Example 1, we can apply the same model building techniques while accounting for the nesting of students within schools, beginning with the straw man model, defined as:

$$\begin{cases} MathScore_{tij} = \pi_{0ij} + \pi_{1ij}(Months_{tij}) + e_{tij} \\ \\ \quad \pi_{0ij} = \beta_{00j} + r_{0ij} \\ \quad \pi_{1ij} = \beta_{10j} + r_{1ij} \\ \\ \quad \beta_{00j} = \gamma_{000} + u_{00j} \\ \quad \beta_{10j} = \gamma_{100} + u_{10j} \end{cases} \tag{3.20}$$

The results of this model are, again, largely uninteresting, but they set up a baseline model from which model comparisons can be drawn. We can then quickly expand the model to include the same covariate applied in Example 1, coded to represent the time students changed grade-levels and began taking more difficult assessments. The first model of interest is then defined as:

$$\begin{cases} MathScore_{ti} = \pi_{0i} + \pi_{1i}(Months_{ti}) + \pi_{2i}(Grdlevel_{ti}) + e_{ti} \\ \\ \quad \pi_{0ij} = \beta_{00j} + r_{0ij} \\ \quad \pi_{1ij} = \beta_{10j} + r_{1ij} \\ \quad \pi_{2ij} = \beta_{20j} + r_{2ij} \\ \\ \quad \beta_{00j} = \gamma_{000} + u_{00j} \\ \quad \beta_{10j} = \gamma_{100} + u_{10j} \\ \quad \beta_{20j} = \gamma_{200} + u_{20j} \end{cases} \tag{3.21}$$

The model displayed in Equation 3.21 fit the data significantly better than the model displayed in Equation 3.20, ($\chi^2 = 5022.60$, $df = 7$, $p < .001$). For space consideration, the results

of our sequence of three-level models are reported beginning with the equation displayed in Equation 3.21 (i.e., the model displayed in Equation 3.20 is excluded). The results are reported in two tables, with the fixed effects displayed in Table 3.3, and the random effects and deviance statistics displayed in Table 3.4. The results of Equation 3.21 (Model 1 in Table 3.3) suggest that students began, on average, scoring 32.00 out of 45 total points. Students then gained, on average, 0.78 questions correct per month, while losing an average of 10.87 points with the transition between grade-levels. We can contrast these results with the model displayed in Equation 3.17, which is an equivalent model but does not account for the nesting of students within schools. We can see that, in this case, accounting for the nesting of students within schools only slightly changes the fixed effects estimates. Yet, as displayed in Table 3.3, we can see that the level 2 variance components are reduced quite dramatically in the three level model. Further, student intercepts, slopes, and the effect of the grade-level transition all vary significantly at level 3. Thus, variance that was attributed to students becomes attributed to schools, and the three level model better represents the relations among the data.

Continuing to expand the model, we can include the same predictors to the model at level 2 that we included in Example 1: (a) special education status, (b) sex, (c) English language learner [ELL] status, and (d) free/reduced lunch status [FRL]. However, because we have a third level, we now also have to decide whether the *effect* of these variables varies by school. That is, does being an ELL student in one school mean the same thing as being an ELL in any other school? While theory should, of course, play a role in these decisions, the relation here is largely an empirical question, and was thus specified all level two predictors as randomly varying between schools. Backward elimination was used again, entering all predictor variables into the model simultaneously. The model was thus defined as:

$$
\begin{aligned}
&MathScore_{tij} = \pi_{0ij} + \pi_{1ij}(Months_{ti}) + \pi_{2ij}(Grdlevel_{ti}) + e_{tij} \\
&\quad \pi_{0ij} = \beta_{00j} + \beta_{01j}(SPED) + \beta_{02j}(Female) + \beta_{03j}(ELL) + \beta_{04j}(FRL) + r_{0ij} \\
&\quad \pi_{1ij} = \beta_{10j} + \beta_{11j}(SPED) + \beta_{12j}(Female) + \beta_{13j}(ELL) + \beta_{14j}(FRL) + r_{1ij} \\
&\quad \pi_{2ij} = \beta_{20j} + \beta_{21j}(SPED) + \beta_{22j}(Female) + \beta_{23j}(ELL) + \beta_{24j}(FRL) + r_{2ij} \\
&\quad \beta_{00j} = \gamma_{000} + u_{00j} \\
&\quad \beta_{01j} = \gamma_{010} + u_{01j} \\
&\quad \beta_{02j} = \gamma_{020} + u_{02j} \\
&\quad \beta_{03j} = \gamma_{030} + u_{03j} \\
&\quad \beta_{04j} = \gamma_{040} + u_{04j} \\
&\quad \beta_{10j} = \gamma_{100} + u_{10j} \\
&\quad \beta_{11j} = \gamma_{110} + u_{11j} \\
&\quad \beta_{12j} = \gamma_{120} + u_{12j} \\
&\quad \beta_{13j} = \gamma_{130} + u_{13j} \\
&\quad \beta_{14j} = \gamma_{140} + u_{14j} \\
&\quad \beta_{20j} = \gamma_{200} + u_{20j} \\
&\quad \beta_{21j} = \gamma_{210} + u_{21j} \\
&\quad \beta_{22j} = \gamma_{220} + u_{22j} \\
&\quad \beta_{23j} = \gamma_{230} + u_{23j} \\
&\quad \beta_{24j} = \gamma_{240} + u_{24j}
\end{aligned}
\tag{3.22}
$$

Again, expanding to three level models can quickly become complex, and the model in Equation 3.22 may appear a bit intimidating upon first glance. Conceptually, however, it is not much more complex than the model displayed in Equation 3.18. The only difference between the models is that Equation 3.22 accounts for the nesting of students within schools, and allows each effect to vary randomly between schools. When the number of random effects is large, as in equation 3.22, estimation can often become difficult, particularly when the number of *j* units is small. The number of parameters to be estimated increases dramatically with each random effect because, in addition to estimating the variance of each parameter, a covariance between parameters is estimated. For example, while only 6 random parameters were estimated at level 3 of Equation 3.21 (variance of u_{00j}, u_{10j}, and u_{20j}, and the covariances between u_{00j} and u_{10j}, u_{00j} and u_{20j}, and u_{10j} and u_{20j}), a total of 120 random parameters were estimated at level 3 of Equation 3.22 (variances of each of the 15 parameters, u_{00j} to u_{24j}, and all covariances). Yet, if

the added complexity better models the observed data, it is likely worth the increased computational demands, as we will obtain a better representation of the relations among the data.

For space considerations, the results of Equation 3.22 are excluded in Tables 3.3 and 3.4. The results of Equation 3.22 suggested that the model overall fit significantly better than the model displayed in Equation 3.21, ($\chi^2 = 858.49, df = 126, p < .001$). Many of the level 3 random effects, however, were not significant. As with model 3.21, students' intercepts, slopes, and the effect of making a grade-level transition all varied significantly between schools. ELL students differed significantly in their intercepts and rate of growth between schools, but the effect of ELL on the grade-level transition variable did not differ significantly between schools. The effect of special education status, sex, and FRL did not vary significantly between schools on any predictor variable. Many of the fixed effects were also not significant. An interesting model building decision thus presented itself: Do we fix the random effects to 0, or remove the non-significant fixed effects (thereby eliminating any possibility of a random effect)? Hox (2010) suggests first fixing the random effects to 0, as the fixed effects may become significant when not allowed to vary at the higher level. Following this advice, all non-significant level 3 variance components were fixed to arrive at the following model:

$$
\begin{cases}
MathScore_{tij} = \pi_{0ij} + \pi_{1ij}(Months_{ti}) + \pi_{2ij}(Grdlevel_{ti}) + e_{tij} \\
\quad \pi_{0ij} = \beta_{00j} + \beta_{01j}(SPED) + \beta_{02j}(Female) + \beta_{03j}(ELL) + \beta_{04j}(FRL) + r_{0ij} \\
\quad \pi_{1ij} = \beta_{10j} + \beta_{11j}(SPED) + \beta_{12j}(Female) + \beta_{13j}(ELL) + \beta_{14j}(FRL) + r_{1ij} \\
\quad \pi_{2ij} = \beta_{20j} + \beta_{21j}(SPED) + \beta_{22j}(Female) + \beta_{23j}(ELL) + \beta_{24j}(FRL) + r_{2ij} \\
\quad \beta_{00j} = \gamma_{000} + u_{00j} \\
\quad \beta_{01j} = \gamma_{010} \\
\quad \beta_{02j} = \gamma_{020} \\
\quad \beta_{03j} = \gamma_{030} + u_{03j} \\
\quad \beta_{04j} = \gamma_{040} \\
\quad \beta_{10j} = \gamma_{100} + u_{10j} \\
\quad \beta_{11j} = \gamma_{110} \\
\quad \beta_{12j} = \gamma_{120} \\
\quad \beta_{13j} = \gamma_{130} + u_{13j} \\
\quad \beta_{14j} = \gamma_{140} \\
\quad \beta_{20j} = \gamma_{200} + u_{20j} \\
\quad \beta_{21j} = \gamma_{210} \\
\quad \beta_{22j} = \gamma_{220} \\
\quad \beta_{23j} = \gamma_{230} \\
\quad \beta_{24j} = \gamma_{240}
\end{cases}
\tag{3.23}
$$

Fixing the random effects to 0 again resulted in a significantly better fitting model, (χ^2 = 138.44, *df* = 105, *p* = .016). Interestingly, removing the random parameters at level 3 resulted in a non-significant random parameter at level 2, r_{2ij}. In other words, once the data were more adequately modeled, students no longer differed significantly in the overall rate at which they dropped with the transition between grade-levels. At level 3, however, the parameter remained significant, suggesting that schools – rather than students – play a larger role in the overall rate at which students drop with the grade-level transition. We thus fixed the r_{2ij} term and reran the model before evaluating fixed effects (note that this model, not Equation 3.23, is reported in Tables 3.3 and 3.4). This model suggested that all terms were significant predictors of student intercepts. However, both ELL and FRL students did not differ significantly in their rate of growth, while both female and SPED students did. Finally, ELL students did not differ

significantly in their initial drop occurring with the grade-level transition, while all other students did. Removing the β_{13j}, β_{14j}, and β_{23j} terms, we arrived at our final model for level 2 (Model 3 in Table 3.3 and 3.4):

$$
\begin{aligned}
&MathScore_{tij} = \pi_{0ij} + \pi_{1ij}(Months_{ti}) + \pi_{2ij}(Grdlevel_{ti}) + e_{tij} \\
&\quad \pi_{0ij} = \beta_{00j} + \beta_{01j}(SPED) + \beta_{02j}(Female) + \beta_{03j}(ELL) + \beta_{04j}(FRL) + r_{0ij} \\
&\quad \pi_{1ij} = \beta_{10j} + \beta_{11j}(SPED) + \beta_{12j}(Female) + r_{1ij} \\
&\quad \pi_{2ij} = \beta_{20j} + \beta_{21j}(SPED) + \beta_{22j}(Female) + \beta_{23j}(FRL) \\
&\quad \beta_{00j} = \gamma_{000} + u_{00j} \\
&\quad \beta_{01j} = \gamma_{010} \\
&\quad \beta_{02j} = \gamma_{020} \\
&\quad \beta_{03j} = \gamma_{030} + u_{03j} \\
&\quad \beta_{04j} = \gamma_{040} \\
&\quad \beta_{10j} = \gamma_{100} + u_{10j} \\
&\quad \beta_{11j} = \gamma_{110} \\
&\quad \beta_{12j} = \gamma_{120} \\
&\quad \beta_{20j} = \gamma_{200} + u_{20j} \\
&\quad \beta_{21j} = \gamma_{210} \\
&\quad \beta_{22j} = \gamma_{220} \\
&\quad \beta_{23j} = \gamma_{230}
\end{aligned}
\tag{3.24}
$$

Once our level 2 model was finalized, we began building our level 3 model by entering school-level predictor variables.

Before adding school-level predictors, let's first reflect back to our final model in Example 1, which is displayed in Equation 3.19. Remember, the exact same data were used for both examples, but because we accounted for the nesting of students within schools in Equation 3.23, we find that many more student-level predictors were significant. Further, FRL was a significant predictor of students' slopes in Equation 3.19, but once we account for the nesting of students within schools, we find it becomes non-significant. Our resulting inferences are then drastically changed between the two models, highlighting the importance of adequately modeling the observed relationships.

Returning to our current model, we need to be careful when adding school-level predictors because our model is already quite complex. All variables should make theoretical sense, otherwise the model may become difficult to interpret and we are more likely to capitalize on chance (i.e., peculiarities of our specific sample). For our example, we have created three aggregate school-level demographic variables: the proportion of students (a) in special education, (b) who are English language learners, and (c) eligible for free or reduced price lunch. All variables were entered grand-mean centered. Theoretically, we could imagine each of these variables playing a role in students' initial status, rate of growth, and rate at which they drop with the grade-level transition (i.e., all the level 2 variables). However, HLM also allows us to include school level variables as predictors of the student-level predictor variables. For instance, we may hypothesize that the magnitude of the FRL effect on students' intercept depends upon the proportion of students who are also FRL eligible in the school. Following this logic, the proportion of each student group at the school level were entered as predictors of each student level predictor variable. All school proportion variables were also entered as predictors of the three main effects from level 2. Our level 3 model with predictors at all levels was thus given by:

$$
\begin{aligned}
&MathScore_{tij} = \pi_{0ij} + \pi_{1ij}(Months_{ti}) + \pi_{2ij}(Grdlevel_{ti}) + e_{tij} \\
&\quad \pi_{0ij} = \beta_{00j} + \beta_{01j}(SPED) + \beta_{02j}(Female) + \beta_{03j}(ELL) + \beta_{04j}(FRL) + r_{0ij} \\
&\quad \pi_{1ij} = \beta_{10j} + \beta_{11j}(SPED) + \beta_{12j}(Female) + r_{1ij} \\
&\quad \pi_{2ij} = \beta_{20j} + \beta_{21j}(SPED) + \beta_{22j}(Female) + \beta_{23j}(FRL) \\
&\quad \beta_{00j} = \gamma_{000} + \gamma_{001}(P_{SPED}) + \gamma_{002}(P_{ELL}) + \gamma_{003}(P_{FRL}) + u_{00j} \\
&\quad \beta_{01j} = \gamma_{010} + \gamma_{011}(P_{SPED}) \\
&\quad \beta_{02j} = \gamma_{020} \\
&\quad \beta_{03j} = \gamma_{030} + \gamma_{031}(P_{ELL}) + u_{03j} \\
&\quad \beta_{04j} = \gamma_{040} + \gamma_{041}(P_{FRL}) \\
&\quad \beta_{10j} = \gamma_{100} + \gamma_{101}(P_{SPED}) + \gamma_{102}(P_{ELL}) + \gamma_{103}(P_{FRL}) + u_{10j} \\
&\quad \beta_{11j} = \gamma_{110} + \gamma_{111}(P_{SPED}) \\
&\quad \beta_{12j} = \gamma_{120} \\
&\quad \beta_{20j} = \gamma_{200} + \gamma_{201}(P_{SPED}) + \gamma_{202}(P_{ELL}) + \gamma_{203}(P_{FRL}) + u_{20j} \\
&\quad \beta_{21j} = \gamma_{210} + \gamma_{211}(P_{SPED}) \\
&\quad \beta_{22j} = \gamma_{220} \\
&\quad \beta_{23j} = \gamma_{230} + \gamma_{231}(P_{FRL})
\end{aligned}
\tag{3.25}
$$

where *P* represents a school level proportion.

The model displayed in Equation 3.24 fit the data significantly better than the model displayed in Equation 3.23. As can be seen in Model 4 of Table 3.3, the proportion of special education students within a school did not significantly relate to any predictor variable. The proportion of ELL students in a school was a significant predictor of students' intercept, though no other predictors. Finally, the proportion of students eligible for free or reduced price lunch was a significant predictor of students' intercept, and the magnitude of the drop that occurred with the grade-level transition. For every 10% increase in the proportion of FRL students within the school, students dropped, on average .10(3.32) = 0.33 points more with the grade-level transition. The proportion of students eligible for FRL within a school was also a significant predictor of the student-level FRL predictor of the grade-level transition. That is, for every 10% increase in FRL-eligible students within the school, the effect of FRL for any individual student was reduced by .10(2.10) = .21 points. This is an interesting and important result. Of course,

without digging deeper we cannot say *why* a higher proportion of FRL eligible students in a school led to a reduced FRL effect for any individual student, but we can postulate factors at play. For instance, schools with a high proportion of FRL eligible students may also be Title 1 schools, and therefore receive extra funding to provide additional educational services.

The discussion of three-level models is concluded here, although it is worth noting that continued model refinement could take place, which would likely lead to a better fitting model and universally significant effects. However, if one's theory postulated that the proportion predictors would all have significant effects, it may be worth stopping model building at this stage to explicitly report the non-significant effects.

Table 3.3

Fixed Effects Parameter Estimates: Example 2

Fixed Effects	Model 1		Model 2		Model 3		Model 4	
	Coeff.	SE	Coeff.	SE	Coeff.	SE	Coeff.	SE
Intercept, γ_{000}	32.00	0.37	34.84	0.34	34.80	0.34	34.62	0.30
Prop SPED, γ_{001}							0.04*	3.35
Prop ELL, γ_{002}							14.28	6.09
Prop FRL, γ_{003}							-6.78	1.37
SPED, γ_{010}			-5.08	0.29	-5.11	0.29	-5.08	0.29
Prop SPED, γ_{011}							-0.82*	3.79
Female, γ_{020}			-0.62	0.22	-0.63	0.22	-0.62	0.22
ELL, γ_{030}			-3.66	0.84	-3.00	0.86	-2.00	0.97
Prop ELL, γ_{031}							-24.50*	15.43
FRL, γ_{040}			-3.11	0.24	-3.03	0.23	-2.88	0.23
Prop FRL, γ_{041}							2.30	1.14
Slope, γ_{100}	0.78	0.02	0.71	0.03	0.73	0.02	0.73	0.02
Prop SPED, γ_{101}							0.16*	0.30
Prop ELL, γ_{102}							-0.39*	0.53
Prop FRL, γ_{103}							0.07*	0.10
SPED, γ_{110}			-0.07	0.03	-0.07	0.03	-0.07	0.03
Prop SPED, γ_{111}							0.05*	0.35
Female, γ_{120}			0.13	0.02	0.13	0.02	0.13	0.02
ELL, γ_{130}			0.06*	0.07				
FRL, γ_{140}			0.04*	0.02				
Transition, γ_{200}	-10.87	0.30	-9.59	0.36	-9.77	0.34	-9.97	0.35
Prop SPED, γ_{201}							-2.88*	4.02
Prop ELL, γ_{202}							1.95*	7.12
Prop FRL, γ_{203}							-3.32	1.50
SPED, γ_{210}			0.80	0.37	0.75	0.37	0.79	0.37
Prop SPED, γ_{211}							1.19*	4.74
Female, γ_{220}			-1.97	0.27	-1.99	0.27	-1.98	0.27
ELL, γ_{230}			0.09*	0.79				
FRL, γ_{240}			-0.98	0.30	-0.56	0.16	-0.49	0.17
Prop FRL, γ_{241}							2.10	0.83

Note. Transition = grade-level transition with new, more difficult measures. Coeff. = Model Coefficient

*coefficient is *not* significant, $p > .05$; All other values significant.

Model 1 = Equation 3.21
Model 2 = Similar model to Equation 3.23, but with r_{2ij} fixed to 0.
Model 3 = Equation 3.24
Model 4 = Equation 3.25

Table 3.4

Random Effects Parameter Estimates: Example 2

Random effects	Model 1	Model 2	Model 3	Model 4
Within-student, e_{tij}	14.45	14.77	14.79	14.77
Student intercept, r_{0ij}	29.25	21.59	21.53	21.54
Student slope, r_{1ij}	0.05	0.01	0.01	0.01
Student transition, r_{2ij}	4.65			
School intercept, u_{00j}	2.49	3.81	3.80	2.28
ELL intercept, u_{03j}		11.53	12.00	10.02
School slope, u_{10j}	0.13	0.02	0.02	0.02
ELL slope, u_{13j}		0.04		
School transition, u_{20j}	1.87	3.33	3.42	3.01
Model deviance (parameters)	79423.36 (16)	78746.50 (34)	78774.37 (26)	78728.02 (41)

Note. Transition = grade-level transition with new, more difficult measures. Coeff. = Model Coefficient. All parameters significant.

Model 1 = Equation 3.21
Model 2 = Similar model to Equation 3.23, but with r_{2ij} fixed to 0.
Model 3 = Equation 3.24
Model 4 = Equation 3.25

Conclusions

The purpose of this paper was to provide an introduction to HLM, and illustrate its application through a less technical lens. It is hoped that the reader was able to grasp a basic understanding of the flexibility of HLM, and some potential applications, even if not every illustration was fully comprehended. The introduction was, however, just that. The range of potential models fit within an HLM framework is vast, and continues to expand at a tremendously rapid pace. One common extension not discussed in this manuscript, are hierarchical *generalized* linear models (HGLM), when the outcome is dichotomous, ordinal, or multinomial. These are the multilevel extensions of logistic, ordinal, and multinomial regression. For example, one may be interested in student and school level factors affecting high-school dropout (i.e., a dichotomous outcome). Or, similarly, one may posit questions about student and school factors affecting ordinal *levels* of achievement (e.g., novice, basic, master, expert). These research questions could be addressed through the HGLM framework.

Other models not discussed in this manuscript include cross-classified models (discussed briefly in Chapter 3), multilevel models for latent variables, and hierarchical multivariate linear models (HMLM). For the interested reader wishing to delve into these topics in greater detail, I recommend Raudenbush and Bryk (2002) and Hox (2010), although a number of other terrific resources abound. For a more in-depth discussion regarding growth models and longitudinal data analysis, I highly recommend Singer and Willett (2003). In the end, the analyst employing HLM is primarily limited only by his or her imagination. Yet, one must also always balance modeling the observed relationships among the data with building parsimonious models that *adequately* represent the observed relationship. Model complexity for the sake of model complexity leads to difficult to interpret results, and, often, overfitting to the sample-specific data. Theory, along

with the empirical data, should always guide model-building decisions. The primary strength of HLM is that it allows analysts to model relationships that are, theoretically and empirically, complex.

References

Anderson, D., Patarapichayatham, C., & Nese, J. F. T. (2013). *Basic Concepts of Structural Equation Modeling* (Technical Report No. 1306). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*, 227-268.

Cronbach, L. J., & Furby, L. (1970). How we should measure "change" - or should we? *Psychological Bulletin, 74*, 68-80. doi: 10.1037/h0029382

Fryer, R. G., & Levitt, S. D. (2004). Understanding the Black-White test score gap in the first two years of school. *The Review of Economics and Statistics, 86*, 447-464. doi: 10.1162/003465304323031049

Grady, M. W., & Beretvas, S. N. (2010). Incorporating student mobility in achievement growth modeling: A cross-classified multiple membership growth curve model. *Multivariate Behavioral Research, 45*, 393-419. doi: 10.1080/00273171.2010.483390

Hannaway, J., & Talbert, J. E. (1993). Bringing context into effective schools research: Urban-suburban differences. *The Journal of Leadership for Effective & Equitable Organizations, 29*, 164-186. doi: 10.1177/0013161X93029002004

Harris, C. W. (1962). *Problems in measuring change. Proceedings of a conference sponsored by tthe Committee on Personality Development in Youth of the Social Science Research Council*. Madison: University of Wisconsin.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. *Educational Psychologist, 35*, 125-141. doi: 10.1207/S15326985EP3502_6

Nese, J. F. T., Biancarosa, G., Anderson, D., Lai, C. F., & Tindal, G. (2011). Within-year oral reading fluency with CBM: A comparison of models. *Reading and Writing: An Interdisciplinary Journal*, 1-29. doi: 10.1007/s11145-011-9304-0

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

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *American Educational Research Journal, 46*, 853-891. doi: 10.3102/0002831209333184

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modeling. *Learning Disabilities: A Contemporary Journal, 2*, 30-38.

Roberts, J. K. (2007). *Group dependency in the presence of small intraclass correlation coefficients: An argument in favor of NOT interpreting the ICC*. Paper presented at the American Educational Research Association.

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

Snijders, T. A. B., & Bosker, R. (2011). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling* (2nd ed.). Thousand Oaks, CA: Sage.

Willett, J. B. (1989). Some results on reliability for the longitudinal measurement of change: Implications for the designs of studies of individual growth. *Educational and Psychological Measurement, 49*, 587-602. doi: 10.1177/001316448904900309

Zimmerman, D. W., & Williams, R. H. (1998). Reliability of gain scores under realistic assumptions about properties of pre-test and post-test scores. *British Journal of Mathematical and Statistical Psychology, 51*(2), 343-351. doi: 10.1111/j.2044-8317.1998.tb00685.x