Chapter 11

MULTILEVEL MODELING FOR PSYCHOLOGISTS

John B. Nezlek

Multilevel analyses have become increasingly common in psychological research, although unfortunately, many researchers' understanding of multilevel analysis has lagged behind this increased interest and use. Many researchers have heard of and are curious about multilevel modeling (MLM), but they are unfamiliar with it, perhaps so unfamiliar that they do not know where to start. This unfamiliarity is probably due in part to the fact that many graduate programs in psychology do not offer (or have not offered) courses in multilevel analysis. This chapter is an attempt to meet this need by familiarizing readers with MLM as it pertains to psychological research broadly defined.

In writing this chapter, I had two goals in mind. First, I wanted readers to learn the basics of multilevel analysis. Second, I wanted to increase readers' awareness of the multilevel perspective so that they might recognize the multilevel features of the data they have collected and would be able to be formulate more clearly research questions that might involve multilevel data. As Kreft and de Leeuw (1998) noted, "Once you know that hierarchies exist you see them everywhere" (p. 1). Conversely, if you do not know how to conceptualize a multilevel data structure and the accompanying analyses, you may not see or recognize hierarchies anywhere.

In this chapter, I provide a rationale for MLM: why it is necessary, its advantages over other techniques, and so forth. I describe the basic structure of univariate multilevel analyses: the nature of the models and the types of parameters they can estimate and how to conduct multilevel analyses, including different aspects of analyses such as centering, modeling error, weighted analyses, and categorical independent and dependent measures. I also offer suggestions about how to interpret the results of analyses and how to report results in papers. Finally, although they are in flux, I discuss software options.

This chapter is intended as an introduction for those who are not familiar with MLM. When writing this chapter, the only statistical training I assumed readers would have was an understanding of basic ordinary least squares (OLS) regression. Analysts who are familiar with the basics of MLM may find some value in my treatment, but advanced topics are not covered. Other chapters in this handbook cover some of these topics, such as Chapters 18, 19, and 20 of this volume.

DEFINING MULTILEVEL ANALYSIS

What Is Meant by *Multilevel*?

A multilevel data structure is one in which observations at one level of analysis are nested (or clustered or grouped) within observations at another level of analysis. Sometimes, multilevel data structures are described simply as *nested* or as *hierarchically nested*. The critical, defining feature of such multilevel data is that observations at one level of analysis are not independent of each other—there is an interdependence among the data that needs to be taken into account. In this chapter, I focus on two-level data structures, but the framework and logic I use to describe two-level structures are readily applicable

DOI: 10.1037/13621-011
APA Handbook of Research Methods in Psychology: Vol. 3. Data Analysis and Research Publication, H. Cooper (Editor-in-Chief)

to data structures with more than two levels. I discuss how to conceptualize the levels of a model in a separate section later in this chapter.

The lack of independence among observations is exemplified in studies of groups (e.g., students in classes or workers in work groups). Individuals in the same group all share the characteristics associated with their group, whereas they differ from each other in terms of individual-level characteristics. In addition, group-level characteristics such as teacher experience or the style of group leaders are likely to vary across groups. Therefore, individual differences such as performance can be examined in terms of explanatory variables at two levels of analysis—the individual (skill, motivation, etc.) and the group (e.g., teacher characteristics)—and members of different groups may vary in terms of measures at both levels of analysis.

The question is how to disentangle relationships between an outcome of interest and measures at multiple levels of analysis. For example, why are Mary's test scores higher than Jane's? Is this because Mary is smarter or works harder than Jane (an individual-level relationship), or is it because Mary's teacher is better than Jane's and so the students in Mary's class tend to have higher grades on average (a group-level relationship), or is it both? Moreover, it is possible that the individual-level relationship between how intelligent a student is and her grades varies across classes. The relationship may be stronger in some classes than in others, and it may be of interest to understand (model) the differences between classes in such relationships.

When addressing such questions it is critical to recognize that relationships at different levels of analysis are mathematically independent. In a study in which persons are nested within groups, relationships at the between-group level tell us nothing about relationships at the within-group level, and vice versa. This independence is illustrated by the data in Tables 11.1 and 11.2. In Table 11.1, the within-group relationships are positive and the between-group relationships are negative, and in Table 11.2, the within-group relationships are negative and the between-group relationships are positive.

TABLE 11.1

Relationships: Positive at Within-Group Level and Negative at Between-Group Level

	Group 1		Group 2		Group 3	
	X	*Y*	*X*	*Y*	*X*	*Y*
	26	31	29	29	31	26
	27	32	30	30	32	27
	28	33	31	31	33	28
	29	34	32	32	34	29
	30	35	33	33	35	30
Mean	28	33	31	31	33	28

TABLE 11.2

Relationships: Negative at Within-Group Level and Positive at Between-Group Level

	Group 1		Group 2		Group 3	
	X	*Y*	*X*	*Y*	*X*	*Y*
	11	18	19	23	19	28
	12	17	20	22	20	27
	13	16	21	21	21	26
	14	15	22	20	22	25
	15	14	23	19	23	24
Mean	13	16	21	21	21	26

Moreover, these two examples do not exhaust the possible combinations. Relationships at the within-group level might vary across groups; some could be positive and some could be negative. Seeing data such as these, one might ask, "Which one is correct?" The answer is that neither is correct. If an analyst is interested in constructs defined at the between-group level, then the between-group relationships are correct. Correspondingly, if the interest is in constructs defined at the within-group level, then the within-group relationships are correct, with the caveat that these relationships may not be consistent across groups.

In MLM the term *group* refers to an organizing unit or cluster. For studies of actual groups (e.g., work groups) this creates no confusion; however, for other types of nesting (e.g., observations within person, such as in a diary study), the term *group* can be confusing. So when observations are nested within people, a person constitutes a group. In a cross-cultural study, cultures might be the groups; in a community psychology study, communities might be the groups, and so forth. Using the term *group* in this way is a tradition in MLM, and as confusing as it may be for those unfamiliar with MLM, I follow this tradition.

Analytic Strategies for Analyzing Multilevel Data

Before describing multilevel random coefficient modeling, which is the currently accepted "gold standard" for analyzing multilevel data sets, I briefly review other methods, in part to highlight some of the strengths of the MLM techniques I consider in detail. One way to distinguish such methods is to distinguish *aggregation* versus *disaggregation* methods. In aggregation methods, within-group summary statistics are calculated (e.g., means) and then analyzed. For example, a researcher might calculate for each U.S. state means for two variables (literacy and percentage of immigrants) and then calculate a correlation at the state level. Such relationships are perfectly acceptable providing one does not commit what is called the *ecological fallacy* (Robinson, 1950), which occurs when it is assumed that the between-group relationships exist at the within-group level. Using the 1930 U.S. Census data, in a classic paper, Robinson found a positive correlation between literacy and percentage of residents who were foreign born at the between-state level but found that the within-state relationship was negative. The panels of data presented in Tables 11.1 and 11.2 illustrate the potential for such ecological fallacies.

In disaggregation analyses, analyses are done at only Level 1 (e.g., the individual level when persons are nested within groups), and relationships between outcomes and Level-2 measures are examined by assigning Level-2 measures to the corresponding Level-1 observations. For example, in a study of work groups, a group-level measure such as the leadership style of the group leader would be assigned to each member of a group. In such analyses, a least-squared dummy variable (LSDV; e.g., Cohen & Cohen, 1983) is often used to control for Level-2 differences in Level-1 measures. In LSDV analyses, a set of $k - 1$ dummy variables are added to the model, where k is the number of Level-2 units. Although such analyses achieve this control, they are fundamentally flawed in at least two ways. First, they assume that relationships between the outcome and the predictors are identical in all groups, an untenable assumption, and second, they do not model error properly. Even if interaction terms between the dummy variables and the predictors are included (which can be unwieldy with many groups and multiple predictors), the two sources of error are not modeled properly.

In a two-level multilevel data structure, there are two sources of error: one associated with sampling Level-1 observations, and the other associated with sampling Level-2 observations. For example, in a diary study in which observations are nested within persons, there is error associated with sampling persons and with sampling days. The error associated with sampling persons is well understood, but a coefficient representing the within-person relationship between daily stress and daily anxiety for an individual also has a sampling error. A coefficient that is based on the 2 weeks a study was conducted will probably be similar to, but not the same as, a coefficient based on another 2-week period. Moreover, the unreliability of such a coefficient (how much it might vary across a series of two periods) needs to be incorporated into significance tests of between-person effects. Studies in which people are nested within groups can be understood in the same way. Most analysts would recognize that students in classes constitute a sample that is meant to represent the population of students. In parallel, the classes in which those students are nested also need to be considered as samples representing the population of classes. It is including these two sources of error simultaneously (the Level-1 error and the Level-2 error) that renders OLS analyses inappropriate. Because of the mathematics involved, OLS analyses can estimate only one error term at a time. Estimating multiple unknowns (simultaneous errors) requires maximum-likelihood estimators, which are the basis for the algorithms used by all MLM programs.

For multilevel data, the maximum-likelihood-based procedures that I discuss in this chapter provide more accurate parameter estimates than comparable OLS analyses, such as using OLS regression to estimate coefficients for individual Level-2 units of analysis and using those coefficients in a single-level analysis between Level-2 units. The greater accuracy of MLM using maximum-likelihood estimators is not hypothetical. It is a demonstrated fact on the basis of the results of Monte Carlo studies in which samples have been drawn from populations with known parameters. The sample statistics that are based on maximum-likelihood procedures are more accurate estimates of such population parameters than the sample statistic based on comparable OLS analyses.

A more detailed discussion of the statistical background of MLM is well beyond the scope of this chapter. A somewhat more detailed discussion of the rationale for MLM can be found in Nezlek (2001), but the truly curious reader is advised to consult a formal text such as Goldstein (2003) or Raudenbush and Bryk (2002). A list of suggested readings is presented at the end of this chapter.

The Basic Models

Consistent with the explanatory framework initially offered by Bryk and Raudenbush (1992), I present the equations for each level of a model separately. Nevertheless, all coefficients at all levels of analysis are estimated simultaneously, so the underlying model is represented by an equation in which the outcome (y) is predicted by the intercepts at each level of analysis, the predictors that are included at each level, and the error terms.

In the standard nomenclature, Level-1 coefficients are represented with βs, (subscripted 0 for the intercept, 1 for the first coefficient, 2 for the second, etc.), and the basic Level-1 model is as follows:

$$y_{ij} = \beta_{0j} + r_{ij}. \quad (1)$$

In this model, there are i Level-1 observations for j Level-2 groups of a continuous variable y. The Level-1 observations are modeled as a function of the intercept for each group (β_{0j}, the mean of y in group j) and error (r_{ij}, which is the deviation of each score in a group from the group mean), and the variance of r_{ij} is the Level-1 error variance.

Each Level-1 coefficient is then modeled at Level 2, and Level-2 coefficients are represented by γs. There is a separate Level-2 equation for each Level-1 coefficient. The basic Level-2 model is as follows:

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

In this equation, the mean of y for each of j Level-2 units of analysis (β_{0j}) is modeled as a function of the grand mean (γ_{00} – the mean of means) and error (μ_{0j}), and the variance of μ_{0j} is the Level-2 variance. When these two basic models are combined, this is referred to as totally unconditional or null because there are no predictors at any level of analysis. The value of unconditional models is discussed in the section Building a Model.

Predictors can be added to this basic model at either level of analysis. Assume a study in which students are nested within classes, and the outcome measure is a test score. At the within-class (individual) level, the relationship between test scores and hours of study could be examined with the following model:

$$y_{ij} = \beta_{0j} + \beta_{1j}\ (\text{Study}_{ij}) + r_{ij}. \quad (3)$$

$$\beta_{0j} = \gamma_{00} + \mu_{0j}. \quad (4)$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j}. \quad (5)$$

In this model, the intercept of y (β_{0j}) for each of j Level-2 classes is modeled as a function of the mean intercept (γ_{00}) and error (μ_{0j}), and the slope (β_{1j}) representing the within-class relationship between scores and studying for each of j classes is modeled as a function of the mean slope (γ_{10} – the average relationship across all classes) and error (μ_{1j}).

In MLM, coefficients are tested for significance against zero, and in this model, the significance test of the mean slope (is the mean slope significantly different from zero?) is made at Level 2, via the γ_{10} coefficient. If the γ_{10} coefficient is significantly different from zero, then the null hypothesis is rejected. The intercept is also tested for significance via the γ_{00} coefficient—that is, is the mean intercept significantly different from zero? The meaning of these tests, that is, what the coefficients represent, will vary as a function of the measures themselves, and most important, the meaning of the intercept will vary as a function of how the Level-1 predictors are centered, a topic discussed in a separate section.

In MLM, the random error terms for Level-1 coefficients (the variances of μ_{0j} and μ_{1j}) are also tested for significance, and such significance tests can be used to make decisions about including or excluding random error terms from models. When an error term for a coefficient is included in a model, the coefficient is referred to as a *random coefficient*, and when an error term is not included, the coefficient is referred to as a *fixed coefficient*. This topic is discussed in more detail in the section Modeling Random Error.

Predictors can also be added at Level 2. Continuing the example, at the between-class level, the relationship between test scores and teacher experience could be examined with the following model:

$$y_{ij} = \beta_{0j} + r_{ij}. \tag{6}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{Experience}_j) + \mu_{0j}. \tag{7}$$

In this model, the mean score for a class (the β_{0j} brought up from Level 1) is being modeled as a function of the intercept and the experience of a teacher. If the γ_{01} coefficient is significantly different from zero, then there is a relationship between a teacher's experience and the average score for students in his or her class. Once again, what these Level-2 coefficients represent will vary as a function of how the Level-2 predictors are centered.

Predictors can be added at both levels of analysis simultaneously. Relationships between test scores and hours of study could be examined at the individual level, and in turn, classroom-level differences in these relationships could be modeled at the between-class level as a function of teacher experience. Analyses examining such relationships are sometimes called *slopes-as-outcomes* analyses because a slope from a lower level (e.g., Level 1) becomes an outcome at an upper level (e.g., Level 2).

$$y_{ij} = \beta_{0j} + \beta_{1j}\,(\text{Study}_{ij}) + r_{ij}. \tag{8}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{Experience}_j) + \mu_{0j}. \tag{9}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}\,(\text{Experience}_j) + \mu_{1j}. \tag{10}$$

In this model, the slope for each class (β_{1j}) is brought up from Level 1 and is modeled as a function of the intercept and the experience of a teacher. If the γ_{11} coefficient is significantly different from zero, then the relationship between test scores and studying varies as a function of teacher experience. Note that *Experience* is included in both Level-2 equations, a topic discussed in the section Building a Model.

Although there is no absolute standard nomenclature, traditionally, in a two-level model, Level-1 coefficients are represented by βs and are analyzed at Level 2 as γs. In a three-level model, Level-1 coefficients are represented by πs and then βs and γs for Levels 2 and 3, respectively. In the popular HLM program, the nomenclature was recently changed to distinguish models in which persons are nested within groups from models in which observations are nested within persons. The traditional βs and γs are used when people are nested within groups, but πs and βs are used when observations are nested within people. (In these systems, people are always represented by βs.) The models and the results of analyses do not vary as a function of which sets of letters are used. The distinction is purely terminological.

Similar to OLS regression, these multilevel models are simply templates that can be applied to various types of data structure. In a study of therapeutic outcomes, clients could be nested within therapists or clinics. In diary-style studies, observations (days or certain types of events such as social interactions) could be nested within persons. In studies relying on reaction times, responses can be treated as nested within persons and experimental conditions can be modeled at the person level. Such applications are limited only by the insight of researchers and their ability to collect the necessary data.

CONDUCTING MULTILEVEL ANALYSES

Building a Model

The first model that should be run is a model that has no predictors at any level of analysis. These are called *unconditional* or *null* models, and in some cases, *variance component* models. Such unconditional models provide the basic descriptive statistics for a multilevel analysis. Although unconditional models typically do not test hypotheses, they provide valuable information about how the total variance of a measure is distributed across the levels of a model. Understanding the distribution of variance can also provide some ideas about how productive analyses at different levels of a model might be. For example, if most of the variance of a measure is at Level 1, it may be difficult to analyze differences in Level-1 means (intercepts) at Level 2. The fact that it may be difficult does not imply that it will not be possible. Small amounts of variance at any level of analysis may still provide a sufficient basis for further analyses at that level. Most multilevel modelers agree that Level-1 models should be finalized before Level-2 differences in Level-1 coefficients are examined. In this instance, *finalized* refers to the selection of predictors and the specification of the error structure. Specifying error structures is discussed in the next section.

Another important recommendation for model building is to forward step rather than backward step models, particularly at Level 1. *Forward stepping* refers to a process that begins with the simplest model to which predictors are added one by one (or in small numbers), with tests of significance at each step. Predictors that are not statistically significant are removed from the model before new predictors are added. *Backward stepping* refers to a process in which all possible predictors are added at the outset and predictors that are not statistically significant are removed sequentially.

Although backward-stepping procedures may be fairly common in OLS regression analyses, because MLM analyses estimate more parameters than seemingly comparable OLS regressions, backward-stepping procedures may stretch what statisticians refer to as the *carrying capacity* of the data. In MLM, the number of parameters that are estimated increases nonlinearly as a function of the number of predictors. For example, in the basic Level-1 model, $y_{ij} = \beta_{0j} + \mathrm{r}_{ij}$, three parameters are estimated: a Level-1 variance, and fixed and random effect for the intercept. If a predictor is added, $y_{ij} = \beta_{0j} + \beta_{1j}\ (x) + \mathrm{r}_{ij}$, six parameters are estimated: the Level-1 variance, a fixed and random effect for both the intercept and the slope (four parameters), and the covariance between the two random effects (one parameter). If a second predictor is added, $y_{ij} = \beta_{0j} + \beta_{1j}\ (x_1) + \beta_{2j}\ (x_2) + r_{ij}$, 10 parameters are estimated: the Level-1 variance, a fixed and random effect for the intercept and the two slopes (six parameters), and the covariances between the three random effects (three parameters).

When adding predictors at Level 2, the norm is to have (initially) the same Level-2 predictors for each Level-1 coefficient. For example, if a Level-1 model had two predictors, $y_{ij} = \beta_{0j} + \beta_{1j}\ (x_1) + \beta_{2j}\ (x_2) + r_{ij}$, then three coefficients, the intercept and two slopes, would be brought up to Level 2. If a Level-2 variable, Z, is used to model the intercept, $\beta_{0j} = \gamma_{00} + \gamma_{01}\ (Z) + \mu_{0j}$, then the other coefficients (the two slopes) should also be modeled as a function of Z, for example, $\beta_{1j} = \gamma_{10} + \gamma_{11}\ (Z) + \mu_{1j}$ and $\beta_{2j} = \gamma_{20} + \gamma_{21}\ (Z) + \mu_{2j}$. One reason for doing this is that if Z is not included as a predictor for a coefficient, it is assumed that there is no relationship between that coefficient and Z. Because all coefficients in a model are being estimated simultaneously (including the covariances between coefficients), the failure to include a relationship between Z and a Level-1 coefficient may lead to a misspecified model.

This discussion of model building has focused on the technical aspects of MLM. Of course, exactly how an analyst chooses to build a model needs to reflect the substantive questions at hand. Nevertheless, these guidelines reflect the knowledge and experience of accomplished multilevel modelers (e.g., Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002).

Modeling Random Error

As discussed, for each coefficient in a model, MLM can estimate a fixed and a random error term. The fixed term is the focus of most hypotheses. It is an estimate of the relationship between a predictor and the outcome. Is the coefficient significantly different from zero? The random error term reflects the ability of the algorithm to separate true and random error, and a coefficient for which a reliable random error term can be estimated is described as *randomly varying*. Although random errors are usually not the focus of hypotheses per se, it is important to estimate random errors properly because in MLM all coefficients are estimated simultaneously. An improperly specified random error structure can lead to inaccurate estimates of the fixed effects, or in more formal terms, a misspecified model.

For those not familiar with techniques such as structural equation modeling (SEM), in which error structures need to be described by the analyst, specifying error terms in MLM may be challenging. In OLS analyses, there is one error term in a model, whereas in MLM, each Level-1 coefficient can have its own error term. In a two-level model with a single predictor at Level 1, there is an error term for both the intercept and the slope. The simultaneous estimation of such errors is part of the reason MLM provides more accurate parameter estimates and significance tests of coefficients than comparable OLS techniques.

Researchers vary in terms of how much they model error structures, and the guidelines I discuss here may not be normative in all disciplines. Nevertheless, this description should provide a good starting point. At the basic level, I recommend deleting from a model error terms that are not statistically significant. Statistical significance in this case refers to the results of a test in which the null hypothesis is that the random error is zero. For most statistical tests, an α level of $p < .05$ is used, but for decisions about the inclusion or exclusion of error terms, most modelers recommend a more relaxed standard such as $p < .10$. This more relaxed standard reflects the fact that, in most cases, the coefficients are theoretically random, and they should be modeled as such if possible. If a random error term cannot be estimated at all reliably to be different from zero ($p > .20$), then it should be deleted from the model so that the information contained in the data can be used to estimate other parameters.

When a random error term is not modeled for a coefficient, the coefficient is described as having been *fixed*. Most important, analysts need to determine whether the inclusion or exclusion of a random error term makes a difference in the coefficients that are the focus of the model—which is usually the fixed effects. Perhaps most important, such differences can include changes in the results of significance tests of the fixed effects. Finally, the addition of Level-2 variables may change the statistical significance of random error terms of the coefficients being brought up from Level 1, and when this happens, analysts need to determine why this occurred. For example, including a Level-2 predictor for a Level-1 slope may account for sufficient variance in that slope to render nonsignificant a random error term that was statistically significant without the Level-2 predictor.

There is considerable confusion about what a nonsignificant random error term means. In the truest sense, it means that there is not sufficient information in the data to separate true (fixed) variability from random variability. It does not mean that a coefficient does not vary at all. It means that the random variability cannot be modeled. This confusion is increased by the fact that when a coefficient is fixed, the estimated (fitted) values for that coefficient will all be the same because no residual estimate has been estimated for that coefficient, and it is such estimates that represent random effects.

Regardless, it is possible (and completely acceptable statistically) to model differences in fixed coefficients. The inability to model random variability does not limit one's ability to model fixed variability. If the random error term associated with a Level-1 coefficient is not significant, the error term can be dropped from the model, but this does not preclude adding a Level-2 predictor to that part of the model. When the variability in a fixed coefficient is modeled in this way, the coefficient is described as *nonrandomly varying*.

Curious readers can conduct two analyses and examine what are called *residual files* to understand this further. The specific content of residual files may vary from program to program, but such files typically contain various estimated values such as fitted values and residual estimates. These data provide a basis for understanding the impact of modeling effects as randomly or nonrandomly varying. As discussed, for fixed coefficients that are not modeled at Level 2, the fitted values will not vary, whereas if a Level-2 predictor is included, the fitted values will vary. The ability to model variability in fixed coefficients reflects the fact that the information provided by the Level-2 predictor provides a basis to allow this. Admittedly, there are differences between nonrandomly varying coefficients and randomly varying coefficients whose variability is modeled at Level 2, but for many purposes, these differences are not relevant.

Centering

Centering refers to the reference value used to estimate a slope for a predictor. For analysts whose primary experience is with OLS regression, centering can be a bit difficult to understand. In most OLS analyses, predictor variables are centered around the sample mean, and the intercept represents the expected value for an observation that is at the sample mean on all the predictors in a model. In contrast, within the multilevel framework, different types of centering options are available, and choosing among these options is far from automatic. A more detailed discussion of this topic and recommendations for choosing among different options can be found in Enders and Tofighi (2007).

At the upper level of analysis in a model (e.g., Level 2 in a two-level model and Level 3 in a three-level model), predictors can be centered in one of two ways: grand mean centered and zero centered (sometimes referred to as *uncentered*). Grand mean centering is similar to the type of centering represented by the standardized coefficients in most OLS regression analyses. Coefficients reflect relationships on the basis of deviations from the total sample means of predictors, and the intercept represents the expected value for an observation at the total sample mean of a predictor or set of predictors.

When a Level-2 predictor is entered uncentered, coefficients reflect relationships on the basis of deviations from scores of zero on the predictors, and the intercept represents the expected value for an observation that has a score of zero on a predictor. Returning to the students within classes example, if sex of teacher was included at Level 2 by using a dummy-coded variable *Male* (coded one for men, zero for women, $\beta_{0j} = \gamma_{00} + \gamma_{01}$ (*Male*) + μ_{0j}), and *Male* was entered uncentered, then the Level-2 intercept of intercepts (γ_{00}) would represent the expected average score for a class with a female teacher (i.e., a teacher for which *Male* = 0).

At lower levels of analysis (e.g., Level 1 in a two-level model and Levels 1 and 2 in a three-level model), there is a third option: predictors can be group mean centered (sometimes referred to as *centered within clusters*, or CWC). When a predictor is group mean centered, coefficients reflect relationships on the basis of deviations from the group mean of a predictor, and the intercept represents the expected value for an observation at the group mean of a predictor or set of predictors. In this instance, the term *group* refers to a Level-2 unit of analysis (in a two-level model). Some use the term CWC to reduce the confusion that may occur because Level-2 units may not be groups.

Centering is a critical aspect of a multilevel model, because the meaning of intercepts and slopes can change dramatically as a function of changes in centering. In some senses, centering is more critical for Level-1 predictors than for Level-2 predictors (in a two-level model) because when considering centering at Level 1, the fact that Level-1 coefficients are *passed up* to Level 2 must be kept in mind. For example, assume a study of student achievement (*y*) in which boys and girls are nested within classrooms with the following Level-1 model. In this model, *Male* is a dummy-coded variable coded one for boys and zero for girls:

$$y_{ij} = \beta_{0j} + \beta_{1j}\,(\text{Male}) + r_{ij}. \tag{11}$$

If *Male* is entered as an uncentered predictor, the intercept represents the expected score for girls in a classroom (i.e., when *Male* = 0). If *Male* is entered grand mean centered, the intercept now represents the classroom mean for achievement adjusted for between-class differences in the distribution of boys and girls. If *Male* is entered group mean centered, then the intercept simply represents the mean for each class unadjusted for differences in sex distributions of classes. Enders and Tofighi (2007, p. 138) provided an algebraic explanation of this.

Regarding recommendations for when to use what type of centering, Bryk and Raudenbush (1992) noted that “no single rule covers all cases” (p. 27), so analysts will need to decide how to center predictors on the basis of the data they have and the hypotheses of interest. Nevertheless, a few broad recommendations are possible. Generally, at Level 2, continuous predictors should be entered grand mean centered. If a continuous measure is entered uncentered, then the intercept represents the expected outcome for a Level-2 unit that has a score of zero on a predictor, something that may not make much sense if a scale does not have a zero point. If on the other hand, a Level-2 predictor has a valid zero point (or can be transformed so that it does, for example, standardizing a measure across all Level-2 observations), entering it uncentered makes more sense. Categorical predictors (dummy or contrast–effect codes) should usually be entered uncentered to maintain the interpretability of the intercept. In general, how predictors behave at Level 2 can be thought of in the same way that one would think of predictors within the context of OLS regression.

Deciding how to center at Level 1 (and at Levels 1 and 2 in three-level model) is somewhat more complex because the coefficients from Level 1 will be brought up to Level 2. If an analyst is interested in what would be the multilevel equivalent of conducting a regression analysis for each Level-2 unit and using the resulting coefficients as dependent measures in another analysis, then Level-1 predictors should be group mean centered. Such analyses take out of the model Level-2 differences in Level-1 predictors. For example, in a study of work groups in which individual-level productivity was modeled as a function of individual differences in job satisfaction, group-level differences in job satisfaction would not contribute to parameter estimates.

Some researchers (e.g., Kreft & de Leeuw, 1998) have suggested that when predictors are group mean centered such differences should be reintroduced in the model by including the means of Level-1 predictors as predictors at Level 2. In contrast, others (e.g., Enders & Tofighi, 2007; Raudenbush & Bryk, 2002) have not seen the need to include such means at Level 2. Regardless, it is important to note that when Level-1 predictors are group mean centered, Level-2 differences in these Level-1 predictors are eliminated from the model, whereas when Level-1 predictors are grand mean centered or uncentered, Level-2 differences in Level-1 predictors are part of the model. Group mean centering holds constant Level-2 means in Level-1 predictors. At present, it is difficult to provide a clear recommendation regarding this, although it is worth noting that most analysts in personality and social psychology (the types of analyses with which I am more familiar) do not enter the means of group mean centered predictors at Level 2. Whether this norm is observed in all disciplines is another matter. Regardless, analysts who are concerned about this should run models with and without these means included to determine what impact their inclusion or exclusion has on their models, with particular attention paid to the impact on the substantive questions at hand.

Grand mean centering predictors at Level 1 adjusts the intercept for each group for group-level differences in predictors. At times such adjustments make considerable sense. For example, assume a school administrator wants to reward teacher performance as defined by their students' success on a math test. Further assume that on average boys are better than girls in math and that the number of boys and girls is not equal across classes. Such a combination would mean that teachers who had more boys in their classes would have higher average math scores than teachers who had fewer boys, assuming that all teachers were equally competent (i.e., that teacher characteristics were not related to math achievement). If some type of coded variable representing student sex was entered grand mean centered at Level 1, the average score in each class (the intercept) would then be adjusted for differences between classes in the number of boys and girls, allowing our administrator to have an estimate of student performance that was not confounded by differences in the gender composition of classes.

When Level-1 predictors are grand mean centered, it is incorrect to use reductions in error variances to make judgments about the strength of Level-1 relationships. When Level-1 predictors are grand mean centered, between group (Level-2) variance is introduced into the Level-1 model, meaning that relationships at Level 1 reflect a mix of variances at both levels of analysis. Finally, Level-1 predictors can be entered uncentered causing the intercept to represent the expected value in each group for an observation with a value of zero on the predictor. This is analogous to how centering effects the intercept in Level-2 models.

Interactions

Within the multilevel framework, interactions can occur either between or within levels. Between-level interactions (sometimes referred to as cross-level interactions or cross-level moderation) occur when a relationship at a lower level of analysis varies as a function of a measure at a level of analysis above it. In a diary study, a Level-1 (within-person) relationship might vary as a function of a Level-2 (person level) variable, such as a personality characteristic. Different programs have different procedures to conduct such slopes-as-outcomes analyses, but assuming the same model is being tested, different programs will provide the same results. Such cross-level interactions were discussed in the section on basic models.

Testing interactions among predictors at the top level of a model is pretty much the same as testing interactions within OLS regression, and analysts are advised to consult Aiken and West (1991). In such cases, the dependent measure is simply a coefficient brought up from a lower level of analysis.

Testing interactions among predictors that are all at the same lower level of analysis (e.g., all at Level 1 in a two-level model) is conceptually similar to testing interactions at the top level of analysis, but there are important procedural differences. Consistent with the recommendations of Aiken and West (1991), for categorical predictors, I recommend simply multiplying them and entering the product uncentered into the model. This can help clarify the results because the intercept remains the expected value for an observation with a value of zero on the predictors.

Aiken and West (1991) recommended centering continuous measures before multiplying them to create interaction terms. Consistent with this, my recommendation is to center continuous Level-1 measures *within* their corresponding Level-2 unit. If they are not centered at all, problems can arise (e.g., in the multilevel setting, colinearity among error terms). If they are grand mean centered, then Level-2 variability in the Level-1 predictors is introduced into the Level-1 interaction term.

The procedures for creating such interaction terms varies from program to program. For example, in HLM, within-level interaction terms (within any level of analysis) need to be created outside of the HLM program and read into the data file the program uses for analysis. When interaction terms involving such centered continuous measures are entered into an analysis, I recommend entering them *uncentered* because the centering has already taken place. Moreover, entering variables representing interactions uncentered, when combined with group mean centered predictors, simplifies the calculation of predicted values. An example of testing within-level interactions using these guidelines can be found in Nezlek and Plesko (2003).

Testing within-level interactions within the multilevel framework has not received that much attention in the literature, but the recommendations I provide here are consistent with what many consider to be good practice. Norms about best practice may vary across disciplines, but the issues I address here should provide, at the least, a good starting point.

Model Diagnostics and Model Fits

Occasionally, models will not converge. Similar to SEM and other techniques that estimate solutions, the maximum-likelihood estimation algorithms that are at the heart of MLM programs fit a model (a set of estimated coefficients and parameters) and then change these estimates to improve the fit of the model—to get the model to fit the data more closely. When improvements of a certain size are reached, the algorithm stops. The size of the improvement that stops the algorithm is known as the convergence criterion. Sometimes, this convergence criterion cannot be reached, and an analysis will keep running until a certain number of iterations are reached. Virtually all programs allow the user to specify the convergence criterion and the number of iterations. Although there is no hard and fast rule, solid models will often converge in a few hundred iterations or less.

My experience is that problems with model convergence invariably reflect some type of problem with estimating error terms, most often, error terms that cannot be estimated reliably. What occurs is that the program is trying to make improvements when improvement is not possible, and it gets caught in a loop (sometimes a local minimum). Occasionally (but not commonly), convergence problems are due to very high correlations between error terms. Although each of a pair of error terms might be able to be estimated reliably, if the correlation between them is too high, the algorithm will get stuck. In my experience, such problematic correlations are very high, .98 and higher.

When convergence problems are due to a bad error term, error terms than cannot be estimated reliably can be dropped from the model. The remedy is not so straightforward when convergence problems are due to an inestimable error covariance and both the error terms involved in the problematic covariance are significant. Some programs (e.g., MlwiN) allow the fixing of specific error covariances, and fixing a covariance would solve such a problem. If an analyst does not want to fix a covariance, then one of the error terms creating the problem can be dropped. Deciding which term to drop can be done on the basis of the impact dropping the term has on the model—less impact being more desirable than more impact. At times, such problems can be due to differences in scales (e.g., some scales have very large variances compared with other) or colinearity between scales.

Although the advent of high-speed computing means that even models that require thousands of iterations to converge will run fairly quickly (a matter of minutes at most), analysts may want to terminate an analysis before convergence and examine the output to determine why a model is having problems converging. My experience has been that premature termination has virtually no effect on the fixed effects in a model when convergence problems are due to problems with the error structure. For example, the fixed effects may be virtually identical for models that have run through 500 and 2,500 iterations, and so "early" models can be examined for problems with error structures.

Models may also not work because of problems with the fixed effects. Such problems are not estimation problems per se; rather they are more structural. Similar to OLS regression, if predictors are linearly dependent, a model will not converge. For example, a model will not run if a dependent measure, y, is predicted by x, z, and $x + z$. Most analysts will be experienced enough to avoid problems caused by the colinearity resulting from using linear combinations of variables in an analysis, but analysts may encounter such problems when they retain the intercept in what is meant to be a *zero-intercept* model. When fitting zero-intercept models, analysts need to be certain to delete the intercept—hence the other term for such models, *no-intercept*. Some applications of zero-intercept models can be found in Nezlek (2003, 2007b).

Although MLM analyses provide measures of overall model fit (a deviance statistic), unlike within the SEM tradition, fit indexes do not figure prominently in the evaluation of MLM results. In MLM, the emphasis is less on the overall fit of the model (i.e., how well a model captures all of the hypothesized relationships among a set of measures) than it is on specific coefficients—for example, is the relationship between two Level-1 variables significant, does it vary as a function of a Level-2 variable, and so forth? There are situations in which the fits of different models need to be compared (e.g., to compare error structures), and deviance statistics can be used when this is necessary, but a discussion of such possibilities is well beyond the scope of this chapter. See Raudenbush and Bryk (2002) for a discussion of the questions that can be addressed by comparing the fits of different models.

SELECTED TOPICS

Missing Data

In terms of the practicalities of setting up data files and models, different programs treat missing data in somewhat different fashions. For example, in the program HLM, when creating the system file that is used for analyses (the MDM file) missing data are allowed at Level 1 of two- and three-level models but are not allowed at Level 2 of a two-level model or at Levels 2 and 3 of three-level models. It is possible to include level units of analysis in the system file that have missing data at Level 1, and such units will be excluded from any analysis in which that measure is included. For example, assume a study in which workers are nested within work groups with four measures for each worker. If one of these measures is missing, that worker can be included in analyses that do not include that missing measure. Earlier versions of the HLM program had an option to use all possible pairs of a set of observations, but that option has been removed from more recent versions. In HLM, it is also possible to eliminate from the system file cases that have any missing data, which will ensure that all results are based on exactly the same data. In HLM, at upper levels of a model, missing data are not allowed, and units of analysis that have missing data are not included in the system file.

In contrast, in MLwiN, cases with missing data are excluded on an analysis-by-analysis basis. Cases that have missing values are excluded from an analysis in which the variable that is missing is included. Most important, if a Level-2 unit has a missing value on a variable that is included in a model, all the

Level-1 cases associated with that Level-2 unit are eliminated from the analysis. Analysts are encouraged to determine exactly how the software they are using treat missing data.

Regardless of how a program treats missing data, analysts may want to estimate missing data to maximize the number of observations that are included in an analysis. Such estimation is particularly important when data are missing for units of analysis at upper levels in a model (e.g., at the person level in a two-level analysis of diary data) because when an upper level unit is excluded, all the lower level units underneath that upper level unit are excluded. (For a discussion of estimating missing data, see Chapter 2 of this volume.)

Irrespective of the software being used, it is important to recognize that missing data within the multilevel context are not missing units of analysis, even though within other analytic frameworks missing units of analysis might be treated as missing data. For example, in a diary study in which participants are asked to provide data for 14 days, within the MLM framework, a participant who provided only 10 days of data would not be considered to have 4 days of missing data. Just as classes might have different numbers of students, individuals can have different numbers of diary entries (days, interactions, etc.). The available Level-1 observations are simply nested within the corresponding Level-2 observations.

Perhaps most important in terms of missing data is understanding why observations are missing. For example, are people who provide only 10 of 14 days of data in a diary study different from (in some meaningful way) people who provide 14 days of data? In this specific instance, a person-level (Level-2) variable could be included in a model and it could be determined whether the absence of data was related to coefficients of interest. Handling missing data is a complex topic, and for present purposes, it will need to suffice to recognize that simply because MLM can accommodate differences in the numbers of Level-1 observations that are nested within Level-2 units does not mean that such differences can be ignored.

Standardization

By design, MLM analyses produce *unstandardized* estimates of coefficients. As far as I know, no program has the option to produce standardized estimates. Moreover, although procedures to standardize coefficients have been proposed (e.g., dividing a coefficient by some type of variance estimate), such procedures are probably, at best, at the edge of being justified statistically. Nevertheless, there are ways to reduce the influence on parameter estimates of differences in the variances of measures, which makes coefficient more readily comparable.

Standardizing Level-2 variables is fairly straightforward and puts all continuous Level-2 measures on the same metric. In studies when observations are nested within persons (e.g., diary studies), this could entail standardizing trait measures such as the FFM, and the same type of standardization could be used when Level-2 units are not persons (e.g., clinics, schools, or work groups). Analysts will need to make such decisions about such matters on the basis of what makes sense within their home disciplines. For example, standardizing measures of personality at Level 2 is probably easily understood by personality psychologists who may be accustomed to interpreting regression analyses by estimating predicted values +/−1 standard deviation (*SD*) from the mean. One of the advantages of standardization at Level 2 is that coefficients for Level-2 predictors represent the change in a Level-1 coefficient associated with a 1-*SD* increase in the Level-2 predictor. Another advantage to standardizing Level-2 measures is that differences in the variances of Level-2 predictors do not contribute to significance tests of differences between Level-2 coefficients.

Standardizing Level-1 variables is not quite so straightforward. Simply standardizing in terms of the total population equates the total variances of different predictors; however, it does not equate the distribution of these variances. Moreover, analysts need to be mindful of the fact that when measures are standardized in terms of the total sample, entering a predictor uncentered is equivalent to entering it grand mean centered. Analysts are advised to avoid standardizing Level-1 variables within Level-2 units. For example, if students were nested within schools, it would not be appropriate to standardize scores within each school. The reason for this is that standardizing in this fashion eliminates from the model differences between Level-2 units in Level-1 measures, and such differences can be important sources of information. In general, covariance modelers prefer to work with raw data rather than standardized data because raw data have more information.

Weighted Analyses

Although assigning weights to observations is not common practice for many psychological researchers, it is a necessity for some. If a researcher is interested in making inferences about populations from which nonrandom samples have been intentionally drawn (e.g., certain groups have been intentionally oversampled), the fact that the sample is intentionally nonrepresentative may need to be taken into account. Within the multilevel framework, such a possibility can exist at each level of analysis. For example, in a study of schools, private schools may be oversampled relative to public schools to provide an adequate basis for drawing inferences about private schools. At the individual level, members of minority groups (defined in various ways such as ethnically, those with a specific diagnosis, etc.) may be oversampled to provide a basis for inference. The weights that are assigned to units at one level of analysis have nothing to do with the weights that are assigned to units at another level of analysis. Analysts who want to weight observations will need to specify weights when they analyze their data. Exactly how to do this will vary from program to program, but the results of the analyses from different programs will be the same because this is a well-understood aspect of MLM.

Power Analysis

Despite the growth in popularity of multilevel models, estimating the power of multilevel data structures is still poorly understood. Certainly, the rule that more observations provide more power holds, but questions remain about how many Level-1 and Level-2 units are needed to test different types of hypotheses. This lack of understanding is primarily due to the fact that MLM analyses estimate so many different parameters that it has been difficult to determine how the power to detect each of these parameters varies as function of the design. Moreover, some discussions of power concern the cost-to-benefit ratios associated with increasing observations at Level 1 versus increasing observations at Level 2, and cost is literally measured monetarily. Such discussions may not be particularly valuable to those for whom such monetary considerations are not important. For those interested in more specific recommendations, I recommend Richter (2006) and Scherbaum and Ferreter (2009), who provided cogent summaries of various rules of thumb, with Scherbaum and Ferreter covering the topic more thoroughly than Richter because their article focuses solely on power.

Given the lack of consensus regarding this matter, I offer the following informal recommendations. When thinking of the power of a multilevel design, keep in mind that as a general rule, intercepts are invariably more reliable than slopes, making them easier to model, particularly when there are cross-level effects. Next, start by thinking about power within a comparable single-level design. How many observations would you need to find a medium or small effect? Within this framework, think of the Level-1 slope in terms of a correlation of a certain size—with smaller, less reliable coefficients corresponding to smaller correlations, and larger, more reliable coefficients corresponding to larger correlations. In terms of examining cross-level interactions, note that two criteria are used to evaluate how easy it will be to model differences in slopes—how large a slope is and how reliable it is. It is such complexity that makes it difficult to provide unambiguous guidelines about power. If you are interested solely in Level-1 relationships, the fact that you may have numerous Level-1 observations nested within Level-2 observations will provide a good basis to estimate some parameters, but a lack of Level-2 units will interfere with your ability to estimate the random effects associated with the Level-1 coefficients and may provide weaker tests of fixed effects.

Effect Size Estimation

Often, researchers want to describe their results in terms of what are commonly called *effect sizes*, which, in the OLS framework, are based on variance estimates. For example, a correlation of .5 can also be explained in terms of the fact that two measures share 25% of their variance. Within the OLS framework, estimating effect sizes using such shared variances or reductions in variance from one model to another is well understood and not particularly controversial.

In contrast, within the multilevel context, estimating effect sizes through the use of shared variances or reductions in variance is neither straightforward nor noncontroversial. To provide a context for this, discussion, I will quote Kreft and de Leeuw (1998) who noted that "in general, we suggest not setting too much store by the calculation of RB2 [Level-2 variance] or RW2 [Level-1 variance]" (p. 119). Part of the difficulty in relying on random error terms to estimate effect sizes is that an additional significant predictor can be added to a model and the Level-1 random error term may not change. In unusual cases, it could increase.

Within the OLS framework, such a situation is not possible. For OLS analyses, significance tests are based on reductions in residual variances, and if a predictor is statistically significant, some reduction in residual variance needs to be associated with the inclusion of this predictor in a model. Within the multilevel framework, significance tests of the fixed effects and estimates of random errors are calculated in separate (albeit related) algorithms. So, it is entirely possible for a Level-1 predictor to have a significant fixed effect but whose inclusion in the model is not associated with any decrease in random error.

Despite the possible problems with estimating effect sizes using random error terms, researchers may still want to do so. In such cases, I urge analysts to be cautious and to remain aware of the problems discussed thus far. Moreover, when estimating effect sizes for Level-1 models, predictors should be entered *group mean centered* (e.g., Kreft & de Leeuw, 1998). If predictors are entered grand mean centered or uncentered, Level-2 differences in the predictors will contribute to the Level-1 variance estimates.

Estimating effect sizes in MLM uses calculations that are similar to those used in OLS analyses. A reduction in variance between two models is calculated, and the difference is divided by the variance in the first (presumably larger) model. This needs to be done separately at each level of analysis, and if multiple coefficients are brought up from a lower level of analysis, such estimates need to be made for each coefficient. If a Level-1 coefficient is modeled as fixed (i.e., no random error terms are estimated for it), effect sizes cannot be estimated using variance reductions because there is no variance to reduce.

Using Coefficients Estimated by MLM in Other Analyses

Most MLM programs allow analysts to save the estimated coefficients from analyses, providing the opportunity to use these coefficients in other analyses, for example, a cluster analysis to identify clusters of Level-2 units on the basis of coefficients. Although technically possible, such analyses may not be optimal because when estimated coefficients are used outside of the multilevel framework this does not take advantage of, or take into account, the sampling error at all levels of analysis. There may be instances in which such uses are unavoidable, that is, there may not be a way to examine the hypotheses of interest within the multilevel framework. Nevertheless, analysts are encouraged to find ways to examine their questions of interest within the multilevel framework.

Nonlinear Analyses

Thus far, MLM has been discussed in terms of continuous, linear dependent measures, and continuous, linear measures are probably the most common type of outcome with which psychologists are concerned. Nevertheless, there are many instances in which outcomes of interest are nonlinear. They could be categorical, for example, recidivism (yes or no), or they could be continuous but not normally distributed, for example, count data such as peer nominations of students in classrooms.

Analyses of nonlinear outcomes require special techniques that take into account the fact that such outcomes violate a critical assumption of MLM—the independence of means and variances. For example, the variance of a binomial outcome is Npq, where N is the number of observations, p is the probability of the event, and q is $1 - p$. Other types of nonlinear outcomes (e.g., multinomial outcomes) also violate this assumption. In terms of MLM, this means that the variance of a Level-1 outcome for a Level-2 unit will vary as a function of the mean outcome within each Level-2 unit.

Similar to the need to conduct logistic regression for nonlinear outcomes in single-level data structures, MLM analyses of nonlinear outcomes require techniques that eliminate the relationships between means and variances. The underlying logic of modeling nonlinear outcomes is the same as that for liner outcomes, but the algorithms differ, and the specific algorithms vary as a function of the type of outcome. For example, analyzing a dichotomous outcome requires the following (Bernoulli) model at Level 1:

$$\text{Prob}(y = 1|\beta_{0j}) = \Phi. \tag{12}$$

In this model, a coefficient, representing the probability of y is then converted to an expected log-odds ($\text{Log}[\Phi/(1 - \Phi)]$), and an expected log-odds is estimated for each Level-2 unit. These log-odds are then analyzed at Level 2 just as coefficients are for continuous measures, and similar to the analyses of continuous measures, predictors can be added at all levels of analysis.

Unfinished Business

A chapter such as this cannot cover all aspects of MLM, and I offer brief comments about a few topics that we could not cover in detail. Moderation within the multilevel framework can be understood within the previous discussion of interactions. Mediation is a much more complex topic, and at present, the best source for advice about how to do this is Bauer, Preacher, and Gil (2006). Understanding how to estimate the item-level reliability of measures administered within the multilevel context is poorly understood. Reliability estimates can be wildly inaccurate unless the nested nature of the data is taken into account. For example, when measurement occasions are nested within persons, it is not appropriate to estimate reliabilities on the basis of means aggregated across occasions because this confounds within- and between-person variances. It is also not appropriate to calculate the reliability for each day of a study and then average the reliability coefficients because this assumes that days can be matched across people, when a basic underlying assumption of the model is that days are randomly sampled. The appropriate method is to conduct a multivariate MLM in which the items for a scale are nested within occasions that are nested within people. The reliability of the Level-1 intercept is the item-level reliability. This topic is discussed in Nezlek (2007b).

In MLM, it is possible to compare any coefficients (or sets of coefficients) using what are called tests of fixed effects, which are basically tests of constraints on a model. For example, assume a Level-1 model with two predictors, $y_{ij} = \beta_{0j} + \beta_{1j}\,(x_1) + \beta_{2j}\,(x_2) + r_{ij}$. The strength of the relationship between y and x_1 can be compared with the strength of the $y - x_2$ relationship by examining the impact on the fit of a model of constraining these coefficients (γ_{10} and γ_{20}) to be equal. If the constraint significantly reduces the fit, then the coefficients are not equal. Such comparisons are influenced by the scales (variances) of the predictors, that is, they are not standardized. When predictors have meaningfully different variances, analysts may want to transform them to reduce differences in variances. Such tests can also be used in conjunction with dummy- and contrast-codes combined with different types of centering options to examine differences across the different categories of a categorical predictor. A discussion of a few ways of doing this is provided in Nezlek (2003).

DETERMINING THE MULTILEVEL STRUCTURE OF A DATA SET

How Many Levels?

In most instances, deciding about the multilevel structure of a data set should be fairly straightforward. Studies of students nested within classrooms, or days nested within persons, or clients nested within clinics are all straightforward two-level models. But what if classrooms are also nested within schools, persons are also nested within groups of some kind (e.g., culture), and clinics are also nested within counties? Should each of these be conceptualized as a three-level model? Unfortunately in terms of simplicity's sake, the answer is "perhaps."

There are two important factors that need to be considered when deciding whether to treat observations as nested. First, is there a reason to believe that there is some dependency among observations? For example, does the county in which a clinic is located really matter in terms of how measures might vary or covary? If it does, then county-level effects should be considered. Second, how many units of analysis are there at each level of analysis? When considering this question, it is important to keep in mind that within a multilevel model, each level of analysis represents a sample from a population. If we have clinics nested within counties, before deciding whether to include county as a level of analysis, we need to consider whether the number of counties we have constitutes a sample that can be used to make an inference to the population of counties. Even if county was conceptually a random variable, two counties would not provide a basis to model the random effect of county, whereas 10 might. Such decisions need to be made on a case-by-case basis.

There are ways to examine differences across units of analysis even if there are not enough units to constitute a level of analysis. In some cases, this may mean conducting analyses that are not formal MLM but that do take into account the possibility that relationships between measures vary across units of analysis. For example, if a cross-culturalist has data from 100 people in two cultures, there are not enough cultures to conduct MLM with people nested within cultures, but other types of analyses can be done—see section Other Types of Multilevel Analyses.

In other cases, levels of analysis that one might want to distinguish but for which there are not enough observations can be represented in another level of analysis. For example, in Nezlek et al. (2008), we collected daily diary data for people in four cultural groups. The planned analyses were three-level models, days nested within people who were nested within cultures. Unfortunately, the four cultures we had did not provide a sufficient basis to estimate random effects for culture for the coefficients of interest. In other words, we did not have enough cultures to generalize to the population of cultures. In light of this, culture was represented as an individual-level variable with a series of dummy codes, and we were then able to compare various coefficients across cultures using tests of fixed effects.

The substantive difference between the analyses we did and the planned (three-level) model is that in the two-level model, country was treated as a fixed effect. Technically speaking, this meant that the inference space of our analyses was limited to the four groups from which we obtained data. Although we were able to compare coefficients for these groups, we could not model (i.e., predict) such differences in a formal way. We were able to establish the fact that the cultural groups differed, but we could not explain (statistically) the variability among the groups.

When deciding about the structure of an analysis, it is important to keep in mind that at least two lower level observations are needed for each upper level unit. For example, in a two-level group study, a group needs at least two people, in a diary study, people need to provide at least 2 days of data, and so forth. If an upper level unit of analysis has only one lower level observation nested within it, there is no nesting—there is no way to separate relationships at the different levels of analysis because the sampling is confounded. Level-2 units that have only one Level-1 observation will be included in an analysis, but they will not contribute to estimates of variances. If an analyst has a data set in which a meaningful majority of Level-2 units have only one Level-1 observation, it might be appropriate to consider whether an MLM is appropriate.

Decisions about how many levels of analysis to use often reflect the tension between the law of parsimony (less is more), and the need to account for dependencies among observations. In the previous clinics within counties example, if I had 15 or 20 counties, I would probably nest clinics within counties simply to take into account any dependency that might exist. On the other hand, sometimes more levels provides no advantage. The data presented in Nezlek and Gable (2001) were originally conceptualized as a three-level multivariate MLM (items for different measures nested within days nested within persons), but we presented the results of two-level univariate MLMs because the results of the simpler two-level models were functionally equivalent to the results of the more complicated three-level multivariate models.

It is not possible to provide rules that cover all cases. In most cases, the number of levels that should be used will be obvious. When it is not, researchers will need to make decisions on the basis of previous practice, their knowledge of the subject matter, and perhaps preliminary analyses describing how important it is to take into account different sources (levels) of variance.

At What Level Should a Construct Be Represented?

In most cases, deciding the level of analysis at which a measure should be placed is straightforward. For example, if workers are nested within groups, then worker-level variables such as time on the job would be Level-1 variables, and group-level measures such as group size would be Level-2 variables. In a daily diary study, day-level data such as daily stressors would be Level-1 data, and person-level data such as personality traits would be Level-2 data.

There may be times when assigning a measure to a level of analysis is not so straightforward. For example, in a study in which students are nested within classes, at what level should student sex be included? If classes have both boys and girls, then student sex is a Level-1 variable. In contrast, if classes are sexually segregated, then student sex would be a Level-2 (or classroom-level) variable.

The critical issue is the extent to which a measure varies within a Level-2 unit of analysis. If it does not vary, it is de facto, a Level-2 variable. Although sex is an individual characteristic, if classes are sexually segregated, then for statistical purposes, sex is a classroom-level characteristic, similar to variables measuring the teacher of a class. Just as all the students in a particular classroom have the same teacher (part of the dependency captured by MLM), if all the students in a class are of the same sex, then sex becomes a classroom-level variable.

The situation becomes a bit more complicated when some classes are single sex and some are mixed sex. In such cases, sexual composition can still be used as a Level-2 predictor (e.g., all male, all female, mixed) but not always in combination with a Level-1 variable representing student sex (e.g., a dummy code for males). If sexual composition is coded as a continuous variable at Level 2 (e.g., percent of males), then a Level-1 variable representing student sex can be included in the same model. In contrast, if sexual composition is coded as a categorical variable at Level 2 (e.g., all male or not), then a Level-1 variable representing student sex cannot be included in the same model because it will create a linear dependence between the Level-1 and Level-2 predictors.

Similarly, measures that represent some type of aggregation of Level-1 measures are treated as Level-2 measures. For example, if a measure of group cohesion that is based on the similarity of scores of the individuals within a group is calculated for each group, then such a cohesion measure is a Level-2 measure—it is the same for all members of a group. The measure that served as the basis for the measure of cohesion is still treated as a Level-1 variable because there is within-group variability.

CROSS-CLASSIFIED AND MULTIPLE MEMBERSHIP ANALYSES

So far, this chapter has concerned nested data structures in which the nesting is straightforward and consistent. Students have been treated as nested within a classroom or a school, clients as nested within a therapist or clinic, and so forth. Nevertheless, students can change schools, and clients can change therapists. Within MLM, when the Level-2 unit within which a Level-1 observation is nested changes, this is called multiple membership. In contrast, cross-classification occurs when a Level-1 unit cannot be uniquely classified into two different classification schemes. The classic example of this is when students are treated as nested within schools and schools are treated as nested within neighborhoods, but some schools have children from different neighborhoods. There are modeling techniques that are appropriate for cross-classified and multiple membership data, and the details of how to conduct and interpret such analyses are beyond the scope of an introduction such as this. Interested readers are encouraged to consult Raudenbush and Bryk (2002) and Rasbash, Steele, Browne, and Goldstein (2009) for details.

Nevertheless, when the number of cross-classified or multiple membership cases is very small it may be appropriate to drop such cases to simplify the analyses. Such a procedure should be followed cautiously, however, and disclosed fully in any description of the analyses of the data in question. For example, multiple membership may be a meaningful datum in and of itself. Dissatisfied clients may be more likely to switch therapists than satisfied clients. If such is the case, analyses that did not include multiple membership clients could provide biased parameter estimates. The extent of such a bias would depend on the number of cases that were dropped.

WHEN TO USE AND NOT TO USE MLM: RELYING ON INTRACLASS CORRELATIONS

Among some scholars (particularly it seems, those concerned with organizational psychology and related topics) there is an active debate about when to use MLM on the basis of the intraclass correlations (ICC) for a set of measures. The ICC is a ratio of the between-unit variance (Level-2 variance) to the total variance (Levels 1 and 2 combined). The argument (more or less) is that if there is not enough between-group variance (i.e., the ICC is low) for a measure or set of measures, then the grouped structure of the data can and should be ignored.

My advice regarding when to use multilevel analyses is quite simple and contrasts sharply with this position. Multilevel analyses should be used when a researcher has a multilevel (or nested) data structure of some kind. Full stop. Although apparently sensible, recommendations about when to use MLM on the basis of ICCs are not made on the basis of sound statistical practice or theory.

First, and perhaps foremost, ICCs provide no indication about how relationships between variables might vary across groups. Such a possibility is represented in the data presented in Table 11.3. Assume six groups of individuals, each measured on two variables, *X* and *Y*. In the data presented in the table, the ICC for both measures is zero. There is no between-group variability in either measure, and the mean for both variables is 15 in all groups. If you ignore the nested structure of the data and treat the observations as individual observations, the correlation between *X* and *Y* is zero. Moreover, if you add a dummy-coded variable representing group membership—the LSDV approach described in the section Analytic Strategies for Analyzing Multilevel Data—the estimated relationship is still zero.

TABLE 11.3

Intraclass Correlations and Within-Group Relationships

Group 1		Group 2		Group 3	
X	***Y***	***X***	***Y***	***X***	***Y***
13	17	13	17	13	17
14	16	14	16	14	16
15	15	15	15	15	15
16	14	16	14	16	14
17	13	17	13	17	13
Group 4		**Group 5**		**Group 6**	
X	***Y***	***X***	***Y***	***X***	***Y***
13	13	13	13	13	13
14	14	14	14	14	14
15	15	15	15	15	15
16	16	16	16	16	16
17	17	17	17	17	17

Nonetheless, inspection of these data reveals that the relationship between *X* and *Y* is not zero. It is perfectly negative in Groups 1, 2, and 3, and perfectly positive in Groups 4, 5, and 6. Admittedly, such variability in relationships could be captured by including interaction terms between each of the predictors (we can assume the variable *X* in this case) and each of the dummy variables. Aside from the awkwardness of such procedures (imagine the model generated with a study of 12 groups with three predictors), such analyses are flawed because they do not take into account the sampling error inherent in a study in which units of analysis are sampled from two populations simultaneously—for example, the group and individual levels.

In addition, it is important to keep in mind that ICCs represent ratios of variances. Even when an ICC is low, there may still be meaningful (absolute) variance at the group level for a data set. Finally, what should the cutoff be for deciding when to use MLM: .20, .15, .30? Any cutoff is arbitrary and is difficult to justify statistically. No doubt, recommendations to ignore the grouped structure of a data set when ICCs are low are well intended. Why use a more sophisticated technique such as MLM when a more familiar and more accessible technique such as OLS regression will suffice? Although such advice may have been appropriate at one time, given the growing familiarity with MLM and the increased accessibility of programs that can conduct MLM, researchers should use MLM to analyze their data whenever possible. I discuss in the next section other ways to analyze nested data structures when it is not appropriate to use MLM.

ANALYZING MULTILEVEL DATA STRUCTURES WHEN MLM MAY NOT BE APPROPRIATE

Although I am a strong advocate of using MLM to analyze nested data structures, there are times when data are nested and it is not possible to conduct the types of multilevel analyses this chapter concerns. For example, assume a researcher collects data at the individual level in three cultural groups. Technically, such a data structure would call for a two-level model in which individuals were nested within cultures. Although an MLM might be able to be fitted to the data (i.e., an MLM program might be able to analyze the data), MLM would not be appropriate for such a data set because there are not enough Level-2 observations (cultures). Recall that we considered the fact that in a multilevel data structure, observations are simultaneously sampled from two populations: the population represented by the Level-1 sampling (people in this example), and the population represented by the Level-2 sampling (cultures in this example). Three cultures is simply not enough to provide a reasonable basis for making inferences about differences among the population of cultures. Admittedly, differences among the specific countries involved can be examined as fixed effects, but no generalization to cultures per se can be made. How many observations is enough to provide a basis for making an inference about the population of cultures? It is not possible to set hard and fast rules for such matters, but researchers can rely on their general knowledge of statistics and inference. Aside from studies that rely on intensive repeated measures from restricted samples (e.g., single-case studies), most researchers would probably assume that 10 or so observations would be the minimum.

Aside from commonsense notions about what constitutes a reasonable basis for drawing inferences to a population, another way to tell whether MLM is not appropriate for a particular data set is to consider how well the data can estimate random effects. Assuming that coefficients are theoretically random, if there is not enough information in a data set to estimate any random error terms, then there might not be enough observations to provide the information needed to estimate random effects. For most psychologists, such problems will consist of an insufficient number of Level-2 observations (cultures, classrooms, clinics, etc.). When deciding not to use MLM when using MLM is dictated by the logic of a data structure, researchers should note the following. The inability to estimate random error terms reliably can and should be used as a justification only when the number of cases is small, for example, when the number of Level-2 units is small (certainly fewer than 10). Random error terms (particularly for slopes) may be difficult to estimate even when there are many observations at both levels of analysis, and in such cases, MLM would be appropriate.

If a researcher decides that MLM is not appropriate for a nested data structure, there are reasonable alternatives. The critically important feature of the recommendations I provide for such alternatives is that they allow for the possibility that relationships among Level-1 measures vary across Level-2 units of analysis. Returning to our three culture example, one way of analyzing these data in a single level would be to conduct what is called a *regression by groups* analysis. A regression equation is estimated for each group (each culture in our example), and the similarity of these equations is compared with an F ratio. Alternatively, dummy or contrast codes representing the interaction of culture and various predictors could be entered into an OLS regression. More simply, correlations can be calculated for each culture and compared with a Fisher's r-to-z transform, and means could be compared with a one-way analysis of variance.

Although such procedures can provide significance tests of differences between groups, and significance tests of within-group relationships, it is essential to recognize their limitations. Most important, the inference of such analyses is limited to the specific groups being studied. Assume we have collected data from Spain, Greece, and the United States. A regression by groups analysis would allow us to conclude whether coefficients from the Spanish sample were different from coefficients from the Greek or U.S. samples, and so forth. We could not make any inferences beyond these samples, and we could not model the differences across the samples—country-level differences that might map onto the differences we found between the three groups. Explaining the differences between the cultures could not be done statistically because three cultures would not constitute a sufficient basis for making inferences about cultures in general.

SOME PRACTICAL MATTERS

Preparing Papers for Publication

Norms vary widely about the details of analyses that should be reported, and the following guidelines need to be considered in that light:

1. Structure of the data—The nesting of the data (what was nested in what) should be described explicitly. This description should include the numbers of observations at each Level of analysis, and for lower levels of analysis (e.g., Level 1 in a two-level model) some indication of the distribution of the number of Level-1 observations for Level-2 units (e.g., the *SD*).
2. Centering—The type of centering used for each predictor should be described explicitly. Coefficients (and the relationships they represent) cannot be understood without knowing how predictors were centered.
3. Error terms—The basis used to include or exclude error terms should be described explicitly. A clear justification should be provided if coefficients are fixed on other than statistical grounds. Nonetheless, extended discussions of error structures are often unnecessary. Unless hypotheses explicitly concern or involve some aspects of the error structure, which may more likely be the case with longitudinal data, extended discussion may distract more than it clarifies.
4. Summary statistics—The mean and variance estimates provided by unconditional analyses are the basic descriptive statistics for MLM analyses. These should be provided for both dependent and independent measures to provide a context for readers to understand the results.
5. Model equations—At present, I think the equations representing the models that were run should be presented. Perhaps after more people become more familiar with MLM, this will not be necessary. Moreover, in keeping with Bryk and Raudenbush (1992), I recommend presenting the equations for each level of an analysis separately. This clarifies what was done, particularly for readers who not are modelers.
6. Statistics—I encourage authors to be lean and mean in terms of the statistics they describe in articles and chapters. For example, the significance of the gamma (Level-2) coefficients that are typically the focus of hypotheses in two-level models are tested with an approximate *t* ratio. This *t* ratio is calculated by dividing an estimate of a fixed effect (gamma) by a standard error. This means that there is no reason to present the gamma, the *t* ratio, and the standard error. Any two will do. I recommend the gamma and the *t* ratio, with an accompanying *p* value.
7. Predicted values—Many MLM analysts recommend interpreting results in terms of predicted values. For categorical predictors, estimated values can be calculated for different groups. For continuous predictors, coefficients can be estimated for units +/–1 *SD*. Keep in mind that the *SD* for a Level-1 measure is *not* the *SD* of that measure from a single-level analysis. The Level-1 *SD* of a measure is the square root of the variance as estimated by an unconditional model. Finally, using predicted values can make real the implications of centering.
8. Indexes of model fits and sequential models—Rarely do I see a justification for presenting indexes of model fits. The fixed effects are the focus of most multilevel hypotheses, and fit indexes include both the fixed and random components. Moreover, comparing models that have different fixed effects requires using full (vs. restricted) maximum likelihood estimators,

> and full maximum likelihood estimators are not as accurate as restricted maximum likelihood estimators under many conditions. Sequential comparisons of models frequently provide little information above what is available from final models. When they provide additional insights, they are certainly valuable. When they do not, they distract more than they inform.

Authors should carefully consider just how relevant certain aspects of an analysis are to the substantive questions at hand. Certainly, different disciplines may have different norms (with good reasons) regarding the details of different types of analyses, and authors will need to recognize the importance of presenting the types of details their readers require.

Software Options

The number of programs that can perform MLM has grown meaningfully over the past 10 to 15 years. Different programs provide the same results assuming the same models—including, among other aspects, type of estimation algorithm—are specified. The computational algorithms underlying MLM analyses are well understood, and there is broad agreement about their application. MLM software falls into two broad categories: general-purpose programs that can do all sorts of analyses including MLM (e.g., SAS), and single purpose programs that can do only MLM (e.g., HLM [see Raudenbush, Bryk, & Congdon, 2004]; and MlwiN [see Rasbash, Charlton, Browne, Healy, & Cameron, 2005]).

For those who are unfamiliar with MLM (or with modeling covariances in general), I recommend starting with a single-purpose program such as HLM. This recommendation reflects the fact that setting up models and interpreting the output is generally more straightforward in single-purpose programs than in general-purpose programs. Single-purpose programs were designed to do only MLM and so the user interface is tailored to MLM analyses. HLM is particularly accessible in terms of model set-up and output. In all-purpose programs, the commands for and results of MLM analyses are just one of many possibilities, and so the user interface is not tailored for MLM.

As analysts and their questions become more sophisticated, they may want to consider using MLM modules within a general-purpose program such as SAS. For example, by combining PROC MIXED with other procedures, analysts can perform advanced analyses such as *mixture models* in which similarities among error structures are used as a basis for categorical analyses. Moreover, for analysts interested in error structures that are more complex than the *standard model* (i.e., covariances between all error terms are estimated), programs such as SAS provide more alternatives. Note that MLwiN, a single-purpose program, also provides the opportunity to model some fairly sophisticated error structures.

When discussing software options with analysts, my primary concern is that they fully understand all the parts of the output of their programs. I have spoken with reasonably experienced analysts who have misinterpreted or misunderstood sometimes critical parts of their outputs. Moreover, such confusion seems to be more common with general-purpose programs such as SAS than it is for single-purpose programs such as HLM. Because they involve multiple levels of analysis, multilevel analyses are more complex than single-level analyses, and analysts are advised to proceed cautiously as they add sophisticated options and to master basic techniques before proceeding to more advanced modeling options.

SUGGESTED READINGS

I recommend the following for those who are interested in learning more about the hows and whys of MLM. For overviews, Raudenbush and Bryk (2002), which is the revision of the first edition by Bryk and Raudenbush (1992); Goldstein (2003); Hox (2002); Kreft and de Leeuw (1998); and Snijders and Bosker (1999) all provide good coverage of MLM, with Raudenbush and Bryk and Goldstein being perhaps being the most complete. Moreover, Raudenbush and Bryk was written to accompany the HLM software, so readers can conduct analyses and read about the same analyses in the book. Those who are interested in using SAS to conduct MLM should start with Singer (1998), and those interested in more advanced aspects of random coefficient modeling per se should consult Littell, Milliken, Stroup, and Wolfinger (1996). Moreover, web-based resources about MLM are constantly evolving and expanding. Any decent search engine should point you in the right direction.

In terms of applying MLM to various substantive areas, I have a written a series of articles and chapters that were intended for social and personality psychologists (Nezlek, 2001, 2003, 2007a, 2007b, 2008), and one that is intended for cross-cultural psychologists (Nezlek, 2010). Moreover, detailed, step-by-step descriptions of how to use MLM to analyze these types of data are presented in Nezlek (2011). Although the content of these articles overlaps with each other and with the content of this chapter, each emphasizes different aspects or applications of MLM. A nice discussion of using MLM to analyze reaction time data is provided by Richter (2006), and although Richter discussed MLM in terms of reading comprehension studies, the extension to other substantive areas that rely on reaction times is fairly straightforward. Clinicians (and others) might want to consult Affleck, Zautra, Tennen, and Armeli (1999) for a discussion of the importance of separating relationships at different levels of analysis.

References

Affleck, G., Zautra, A., Tennen, H., & Armeli, S. (1999). Multilevel daily process designs for consulting and clinical psychology: A preface for the perplexed. *Journal of Consulting and Clinical Psychology, 67*, 746–754. doi:10.1037/0022-006X.67.5.746

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

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11*, 142–163. doi:10.1037/1082-989-X.11.2.142

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*, 121–138. doi:10.1037/1082-989X.12.2.121

Goldstein, H. I. (2003). *Multilevel statistical models* (3rd ed.). London, England: Edward Arnold.

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

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

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event and interval contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin*, 27, 771–785. doi:10.1177/0146167201277001

Nezlek, J. B. (2003). Using multilevel random coefficient modeling to analyze social interaction diary data. *Journal of Social and Personal Relationships, 20*, 437–469. doi:10.1177/02654075030204002

Nezlek, J. B. (2007a). A multilevel framework for understanding relationships among traits, states, situations, and behaviors. *European Journal of Personality, 21*, 789–810. doi:10.1002/per.640

Nezlek, J. B. (2007b). Multilevel modeling in research on personality. In R. Robins, R. C. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 502–523). New York, NY: Guilford Press.

Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2*, 842–860. doi:10.1111/j.1751-9004.2007.00059.x

Nezlek, J. B. (2010). Multilevel modeling and cross-cultural research. In D. Matsumoto & A. J. R. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 299–347). Oxford, England: Oxford University Press.

Nezlek, J. B. (2011). *Multilevel modeling for social and personality psychology*. London, England: Sage.

Nezlek, J. B., & Gable, S. L. (2001). Depression as a moderator of relationships between positive daily events and day-to-day psychological adjustment. *Personality and Social Psychology Bulletin*, 27, 1692–1704. doi:10.1177/01461672012712012

Nezlek, J. B., & Plesko, R. M. (2003). Affect- and self-based models of relationships between daily events and daily well-being. *Personality and Social Psychology Bulletin*, 29, 584–596. doi:10.1177/0146167203029005004

Nezlek, J. B., Sorrentino, R. M., Yasunaga, S., Otsubo, Y., Allen, M., Kouhara, S., & Shuper, P. (2008). Cross-cultural differences in reactions to daily events as indicators of cross-cultural differences in self-construction and affect. *Journal of Cross-Cultural Psychology, 39*, 685–702. doi:10.1177/0022022108323785

Rasbash, J., Charlton, C., Browne, W. J., Healy, M., & Cameron, B. (2005). MLwiN version 2.02 [Computer software]. Bristol, England: Centre for Multilevel Modelling, University of Bristol.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009). *A user's guide to MlwiN, Version 2.10*. Bristol, England: Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). HLM 6 for Windows [Computer software]. Lincolnwood, IL: Scientific Software.

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

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*, 221–250.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*, 351–357. doi:10.2307/2087176

Scherbaum, C. M., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods, 12*, 347–367.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*, 323–355.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. London, England: Sage.