CHAPTER 20

Multilevel Modeling Analyses of Diary-Style Data

JOHN B. NEZLEK

Researchers who use ambulatory assessment methods and various types of diaries are increasingly (almost invariably) using some type of multilevel technique to analyze their data. This reflects the fact that the data collected in such studies are inherently multilevel. A sample of individuals provides data on a repeated basis, creating a multilevel data structure in which people constitute one level of analysis and the repeated measures they provide constitute another level, or levels, of analysis. In this chapter I discuss multilevel random coefficient modeling (MLM), the technique that is currently thought to be the best way to analyze such multilevel data structures. I introduce MLM and present ways of using MLM that are well suited for analyzing data generated in ambulatory assessment and diary studies.

When writing this chapter, I tried to address the needs of two audiences: researchers who are quite familiar with MLM to analyze data collected using ambulatory assessment and other intensive repeated measures, and scholars (both new and established) who are not at all familiar with such applications. To address the needs of these different audiences I have provided sufficient introductory material to allow those who are unfamiliar to understand the basic principles involved, while describing more sophisticated applications for the benefit of those who are already familiar. Consequently, those who are familiar with MLM analyses of diary data may wish to skip or skim introductory sections and focus on sections dealing with specific topics or applications.

For those who are not familiar with MLM and want to know more about the technique, I recommend the following. For introductions to MLM per se, Raudenbush and Bryk (2002), Kreft and de Leeuw (1998), Snijders and Bosker (1999), and Hox (2002) are all accessible to the nonexpert (albeit with some extra effort to get through some sections). In terms of using MLM to analyze diary-style data per se, I have written a few articles that may be helpful (Nezlek, 2001, 2003, 2007a, 2007b). Although there is some overlap among them, the articles vary in terms of focus and detail. I also recently published a book that should be helpful (Nezlek, 2011).

Consistent with the focus of this handbook, I discuss MLM in terms of the types of multilevel data that are frequently collected in ambulatory assessment and diary research, although much of what I discuss can be applied to other types of data. Moreover, to illustrate certain points, I often refer to my own research. I have done this not because I am the only the person who has used MLM to analyze these types of data (quite the opposite; there are many experienced and knowledgeable scholars who have published MLM-based studies in this area); rather, I am more familiar with my own studies than I am with the work of others.

Conceptual and Statistical Background

Multilevel data structures are sometimes referred to as *nested data*, because observations at one level of analysis are nested within observations at another level of analysis. Whether the data are what Wheeler and Reis (1991) described as *interval contingent* (e.g., a daily diary study), signal contingent (e.g., a beeper study), or event contingent (e.g., a social interaction diary study) does not matter for present purposes. A study in which people provide the same data each day for a period of time would be described as *days nested within persons.* If multiple assessments are collected each day, it may be useful to think of such data in terms of three levels: (1) assessments nested within days and days nested within people; if people describe their social interactions they had over a period time, such data would be described as interactions nested within persons.

Inferential statistics provide a basis to make inferences about the characteristics (parameters) of the population from which a sample of observations has been drawn. It is critical to recognize that a multilevel data structure is created by drawing samples from different populations, and so there are two (or more) targets of inference. In a daily diary study, one sample is people and the other is days. In a social interaction diary study, one sample is people, and the other is interactions. In a daily diary study, researchers may want to draw inferences about day-level relationships such as that between daily stress and daily mood. They may also want to draw inferences to the population of people about how mean levels of daily observations (e.g., mood) and within-person relationships between daily observations (e.g., mood–stress) vary across persons.

Such multiple sampling means that the sampling error associated with drawing samples at each level of analysis needs to be taken into account. For example, a coefficient calculated over a specific 2-week period representing a within-person relationship between two daily measures is, in fact, sampled from a population of such coefficients. For the same person, a coefficient based on data collected during the first 2 weeks of July will not be exactly the same as a coefficient based on data collected during the first 2 weeks of August, even if that person's life has not changed. When making an inference to the population of days, there is some error in an estimate of a population parameter based on a sample of days, just as there is some error associated with estimates of population parameters based on a sample of persons.

A critical shortcoming of ordinary least squares (OLS) regression analyses of the types of nested data typically collected in diary-style studies is their inability to take into account simultaneously the error (variance) associated with the sampling of observations at multiple levels of analysis. For example, if within-person relationships between variables were represented with within-subject correlations, and these correlations were then used as dependent measures in another analysis, the random error associated with these correlations would not be part of the analysis. Single-level OLS regression analyses in which person-level variables are assigned to daily measures are also flawed for similar reasons, no matter how elaborate the correction or adjustment. These and other shortcomings are described in more detail in Nezlek (2001, 2003, 2007b, 2011).

At this point it suffices to note that the type of multilevel analyses described in this chapter provide the most accurate estimates of parameters for common types of multilevel data structures in ambulatory assessment research. Accuracy in this instance is defined in terms of the results of Monte Carlo studies, in which samples are drawn from a population with known parameters, and the correspondence between these parameters and the estimates produced by different techniques are compared.

Basic Models in the Analyses of Diary-Style Data

Within the nomenclature of MLM, the levels in a multilevel dataset are labeled level 1, level 2, and so forth, with lower numbers nested within higher numbers. For a study in which participants provided data each day for some period of time, daily observations would constitute the level-1 data, and data describing the participants would constitute the level-2 data. Note that the number of level-1 observations does not have to be the same for all level-2 units. People can provide different numbers of days. The minimum number of level-1 observations a level-2 unit should have is two. If there is only one level-1 observation for a level-2 unit, the variance between the two levels cannot be distinguished. In a diary study, if people provided only 1 day of data, it would not be possible to distinguish within- and between-person variance (level 1 and level 2, respectively), because there is no way to estimate the within-person variance. How to conceptualize the levels in an analysis is discussed in more detail later. In this chapter, I focus on two-level models, although the principles and techniques I describe can be applied to models with more than two levels.

Within MLM, level-2 units are often referred to as groups, even though they may not be actual groups. For example, in a diary study in which days are nested within persons, persons are typically described as a grouping variable. This convention may be confusing initially, but it is well established and unlikely to change. When discussing MLM, I rely on the explanatory framework developed by Bryk and Raudenbush (1992). In this framework, they present separate equations for each level of a model. For a diary study, there would be a day-level equation (level 1) and a person-level (level 2) equation. Within this framework, coefficients at lower levels of analysis are "brought up" to higher levels of analysis. Although all coefficients in a MLM are estimated simultaneously, I think separating the equations for each level of analysis clarifies what is being done. Moreover, I think presenting MLM in this way is particularly helpful for readers who may not be familiar with MLM.

The basic two-level model is presented below. In this model i observations of a continuous measure y are nested within j level-2 units. Although conventions vary somewhat, I describe models using β (beta) for level-1 coefficients and γ (gamma) for level-2 coefficients.

$$y_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$

Such a model is referred to as an *unconditional* or *null model*, because there are no predictors at any level. Level-1 observations for each level-2 unit are modeled as a function of the mean for that level-2 unit (β_{0j}), and the level-1 variance is the variance of the deviations of each score from that mean (r_{ij}). In turn, the means for each level-2 unit are modeled as a function of the grand mean (the mean of the means, γ_{00}), and the level-2 variance is the variance of the deviations of each mean from the grand mean (u_{0j}).

If daily observations of mood were nested within person, the level-1 equation would be the within-person (day-level) model, β_{0j} would be the mean mood for each person collapsed across days, and the level-1 variance would be the day-level (within-person) variance. Correspondingly, the level-2 equation would be the between-person model, γ_{00} would be the grand mean for mood, and the level-2 variance would be between-person variance.

To examine level-2 differences in level-1 intercepts (e.g., relationships between Neuroticism and mean daily Mood) a predictor could be added at level 2 (person-level), as shown below. Such an analysis is functionally equivalent to calculating a mean score for Mood for each person, then correlating this mean with Neuroticism scores at the between-person level. If the γ_{01} coefficient is significantly different from 0, then the relationship between mean daily Mood and Neuroticism is statistically significant.

$$y_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01} \text{ (Neuroticism)} + u_{0j}$$

To examine level-1 relationships (e.g., within-person relationships between Stress and daily Mood) predictors can be added at level 1 (the day-level), as shown below.

$$y_{ij} = \beta_{0j} + \beta_{1j} \text{ (Stress)} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Note that for each level-2 unit (e.g., each person) an intercept (β_{0j}) and a slope (β_{1j} representing the relationship between the predictor and the outcome, Stress and Mood) are estimated, and these level-1 coefficients are then analyzed at level 2. Such an analysis is roughly equivalent to calculating a regression equation for each person, then analyzing these coefficients at the person level. The hypothesis that the mean intercept is different from zero is tested with the γ_{00} coefficient, and the hypothesis that the mean slope (the relationship between Stress and Mood) is different from 0 is tested with the γ_{10} coefficient. Note that in this model there are separate error terms for the intercept and the slope, a topic discussed in the next section.

It is also possible to examine how level-1 slopes vary as a function of level-2 measures. Such an analysis is sometimes referred to as a "slopes as outcomes" analysis because a slope from level 1 becomes an outcome at level 2. When a level-1 slope varies as a function of a level-2 measure, this is sometimes called a "cross-level interaction," or a moderated relationship because a relationship at one level of analysis is varying as a function of a variable at another level.

The model below examines whether the Mood–Stress slope at level 1 varies as a function of Neuroticism at level 2. Note that Neuroticism is included in the level-2 equations for both the intercept and the slope. Even if there is no hypothesis concerning the intercept, it is best to include the same predictors in all level-2 equations (at least initially). This recommendation reflects the fact that estimating coefficients in MLM is based on covariance matrices, and if a level-2 predictor is not included in the equation for a level-1 coefficient, then it is assumed that the level-2 predictor is not related to that particular level-1 coefficient. If this assumption is incorrect (i.e., if the predictor is in fact related), and the predictor is not included in the model, the model may be misspecified.

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}\,(\text{Stress}) + r_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\,(\text{Neuroticism}) + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\,(\text{Neuroticism}) + u_{1j} \end{aligned}$$

Random Error and Modeling the Variability of Coefficients

One of the advantages of MLM analyses compared to OLS techniques is the ability to model multiple error terms simultaneously. In the previous example, the intercept and slope each had an error term. When a level-1 coefficient is modeled with an error term at level 2, it is described as *randomly varying*. When a level-1 coefficient is modeled without a random error term, it is described as *fixed* (not to be confused with the estimate of the fixed effect, i.e., the coefficient itself). Finally, if a level-1 coefficient is modeled without a random error term but differences in the coefficient are modeled as a function of a variable at level 2, it is described as *nonrandomly varying*.

Some guidelines for modeling the error structure of a model are presented in the section below on model building. My sense is that most, if not virtually all, coefficients in diary-style research should be modeled as randomly varying if the random effects can be estimated reliably. I have read articles in which authors have not modeled coefficients as *random* (a term used instead of the more cumbersome *randomly varying*), because they believed there was no compelling theoretical reason to assume a coefficient (e.g., a level-1 slope) should vary across level-2 units. While there may not be a reason to expect a coefficient to vary, such an expectation on its own is a poor justification for *assuming* that it does not vary, particularly when such an assumption can be tested. If, as discussed below, the data do not provide a basis to estimate such variability, fine, do not model it. On the other hand, if the data provide a basis for such an estimate, it seems inappropriate to ignore the reality suggested by the data, irrespective of what some theoretical position might lead one to assume.

Regardless, it is possible to model level-2 differences in a level-1 coefficient even when the random error term associated with such a coefficient is not significant. Some might argue that the lack of significant random error term for a coefficient means there is no variability and there is therefore nothing to model. Although such an approach is understandable, the lack of a significant random error term simply means that the data cannot estimate the random error—that is, the data do not provide a sufficient basis for distinguishing fixed (true) and random variation. The lack of a significant random error term does not mean there is no variability: It means that the variability cannot be modeled given the data at hand. When a level-2 predictor is included in a model (when a coefficient is modeled as *nonrandomly varying*), more information is introduced into the model, and this information may provide a basis to model the variability in a coefficient.

Finally, it is important to keep in mind that it may be useful to model the variability in a coefficient even if the fixed effect is not significantly different from 0. For example, in the case of a slope, it is possible that the slopes for half the level-2 units (e.g., people) are positive, and half are negative, resulting in a mean of 0. If this is the case, it may be possible to find a level-2 variable that can model such differences.

Model Building

Although there are no absolute rules about how to build models in MLM analyses, there are guidelines. Analysts should first run unconditional models, described previously, for all the measures in a study. Such models provide the basic descriptive statistics of MLM analyses—estimates of means and variances at each level of analysis. Moreover, these estimates of the distribution of variances can provide some insight into the levels of analysis that might be the most productive. If there is little variance in a measure at a certain level of analysis, it may be difficult to model that variance. For example, in Nezlek, Kafetsios, and Smith (2008), a study in which social interactions were nested within persons, we were able to model person-level (level 2) differences in the positive affect people experienced in their interactions, but we were not able to model person-level differences in the negative affect they experienced. One reason for this difference may have been the relatively small amount of between-person variance in our measures of negative affect. Noting this, even when there is relatively little variance at a level of analysis, it may still be possible to model that variance. The lack of variance at a level of analysis doe not indicate that variance cannot be modeled. It simply means that it may be difficult to model. Finally, the variance estimates provided by unconditional models can be used to calculate intraclass correlations (ICCs).

How to add predictors is another important aspect of model building, and there are two widely recognized guidelines for this. First, level-1 models should be finalized before level-2 predictors are added. Finalization includes specifying the predictors and the error structure. Second, in terms of adding predictors, forward stepping rather than backward stepping procedures are preferred. Unlike many OLS regression analyses in which all possible predictors are added and those that are not significant are dropped from the model, within the multilevel context, it is advisable to enter predictors singly (or a few at a time), evaluate their contribution to the model, then decide to retain or drop them—building a model rather than tearing one down.

In MLM, forward stepping procedures are preferred (particularly at level 1) because of the number of parameters estimated in a model. As predictors are added, the number of parameters that are estimated may tax what is sometimes called the *carrying capacity* of a dataset—the number of parameters a dataset can estimate. For a level-1 model with a single predictor, six parameters are estimated: the level-1 variance, the fixed and random terms for the intercept (2), the fixed and random terms for the slope (2), and the covariance between the two random error terms (1). When there are two level-1 predictors, 10 parameters are estimated: the level-1 variance, a fixed a random term for the intercept and for each of the two predictors (6), and the covariances between the three random terms (3). When there are three predictors, 14 parameters are estimated: a fixed and random term for the intercept and for each of the three predictors (8), and the covariances among the four random terms (6). In comparison, in OLS regression, only one error term is estimated, and only one parameter is estimated for each predictor.

The norm within the multilevel context emphasizes tighter, more parsimonious models that include only variables that have explanatory power over models with many predictors that provide less precise estimates of individual coefficients. Of course, analysts will need to take into account the norms of their home disciplines regarding the type of models considered to be appropriate. Nevertheless, to the extent possible, they should be guided by the law of parsimony—in this regard, less is truly more.

The final consideration I discuss is modeling error. Unless there are compelling reasons to think otherwise, I recommend that coefficients in diary-style research should be modeled as random. As discussed earlier, occasions of measurement are invariably sampled from some population of occasions, and this sampling error needs to be taken into account. Nevertheless, it may not be possible to model reliably the random error for all the coefficients in a model, raising questions about how to treat nonsignificant random error terms. There is broad agreement, although not a true consensus, that nonsignificant random error terms should be deleted from models.

Following such a guideline requires significance tests of random error terms. There seems to be some agreement that a more liberal p-level than .05 should be used for making decisions about retaining random error terms—something like $p < .10$ as a cutoff, with anything greater than .15 or .20 being grounds for dropping a term. What about values in the .10 to .15 range? In such cases, I recommend running the analysis with and without such terms and seeing what impact the inclusion or exclusion has on the fixed effects, which are the focus of most hypotheses. If excluding the terms makes no difference, then they should be dropped. If their presence changes the fixed effects, why this occurred should be considered. Having a clear standard for including or excluding random error terms provides a context for modeling error structures properly. To facilitate this process, I recommend evaluating error structures before examining fixed effects. If this is done, decisions about error terms will be made without regard to the impact changes in the error structure may have on significance tests of the fixed effects.

Centering

A critical aspect of multilevel modeling is the centering of predictors. *Centering* refers to the reference value around which the deviations of a predictor are taken, and there are different centering options available for different levels of analysis. It is critical to keep in mind that centering determines what intercepts represent. In this section, I provide a brief discussion, with some basic recommendations, although as noted by Bryk and Raudenbush (1992, p. 27), "No single rule covers all cases," and analysts need to make choices based upon their specific needs. See Enders and Tofighi (2007) for a thorough and informed discussion of centering options in MLM.

At the top level of a model (level 2 in a two-level model, level 3 in a three-level model, etc.), there are two options. Predictors can be entered either grand-mean-centered or -uncentered (sometimes referred to as zero-centered). The *grand mean* is the mean of a variable for the entire sample. When a predictor is entered grand mean–centered, the slope is calculated based on deviations from the grand mean, and the intercept represents the expected outcome for a unit (e.g., a person) that is at the grand mean of the predictor. This is the same as in OLS regression.

When a predictor is entered uncentered, the slope is calculated based on deviations from 0, and the intercept represents the expected outcome for a unit that has a value of 0 on the predictor. If gender was represented with a dummy-code (0 = male, 1 = female), and gender was entered uncentered, then the intercept would represent the expected value when gender = 0 (i.e., for a man). Although some programs allow zero-centered OLS regression, such use is not that common.

Selecting a centering option at this level is fairly straightforward. Continuous predictors are usually entered grand mean–centered, and categorical predictors can be entered either way. There is some sense among modelers that it is better to enter categorical predictors uncentered because this makes the meaning of the intercept clearer. Moreover, continuous predictors can be entered uncentered if 0 is a meaningful value for the predictor. If 0 is not meaningful (e.g., a predictor is on a 10- to 100-point scale), it makes little sense to enter a predictor uncentered and estimate an intercept that represents an expected value of an outcome when the predictor cannot be 0.

At other levels of analysis (e.g., level 1 in a two-level model), predictors can be entered uncentered, grand mean–centered, or group mean–centered. Moreover, how predictors are centered at lower levels of a model typically has more of an influence on the results of an analysis than how predictors are centered at the top level. This is because coefficients are passed "up" from lower levels to be analyzed at higher levels, and centering changes what these coefficients represent. I discuss centering options in terms of level 1 of a two-level model, but the same principles apply to lower levels per se (e.g., level 2 of a three-level model).

Similar to centering predictors at the top level, when a level-1 predictor is entered uncentered, the intercept (which is now the intercept for a group of observations) represents the expected outcome for an observation in a group that has a value of 0 on a predictor. So, if we had a diary study in which days were nested within persons and a variables DAYS was coded weekend = 0, weekday = 1, if DAYS was entered uncentered, then the intercept for each person would represent the expected score for weekend days (i.e., when the predictor DAYS = 0).

When a level-1 predictor is grand mean–centered, the intercept represents the expected outcome for an observation in a group that is at the grand mean of a predictor (i.e., the mean of all observations in the sample). One of the results of grand mean–centering level-1 predictors is that when predictors are grand mean–centered, level-1 intercepts are adjusted for level-2 differences in predictors. For example, in a social interaction study, a hypothesis may concern the relationship between Neuroticism and the Intimacy people find in their interactions, and the general structure of the analyses is that interactions are nested within persons—as below.

$$y_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{Neuroticism}) + u_{0j}$$

If a variable representing whether an interaction was with a close friend was entered grand mean–centered at level 1, the resulting intercept would represent the mean intimacy people found in their interactions, *adjusted for individual differences in the percent of interactions that involved a close friend.*

$$y_{ij} = \beta_{0j} + \beta_{1j}\ (\text{Close Friend}) + r_{ij}$$

If Neuroticism was also related to the percent of interactions people have with close friends (let's assume negatively), it is possible that relationship tested by the $?_{01}$ coefficient in the first model is confounded. If more neurotic people have fewer interactions with close friends than less neurotic people, and interactions with close friends are more intimate than interactions with non-close others, the level-2 relationships between Neuroticism and mean Intimacy may reflect individual differences in the distribution of interactions. By entering Close Friend grand mean–centered at level 1, such a confound is corrected.

At lower levels of a model, predictors can also be group mean–centered. When a predictor is group mean–centered, deviations are taken from the mean of a predictor for each group, and the intercept represents the expected outcome for an observation for which the predictor is the group mean of the predictor. Aside from rounding error, when predictors are entered group mean–centered, the intercept for each group is unchanged from an unconditional model. Conceptually, entering predictors group mean–centered is similar to conducting a regression analysis for each group and using the coefficients from these analyses as the dependent measures in another analysis, a procedure that is sometimes referred to as *two-stage regression.*

Similar to the recommendations for centering at level 2, there is a general sense that categorical predictors should be entered uncentered, in part because this facilitates interpretation of the intercept—the intercept is now the expected value when a predictor is 0. For continuous predictors at level 1, there is some disagreement about grand- versus group-mean centering. I (and many other analysts) favor group-mean centering continuous predictors. When predictors are group mean–centered, level-2 differences in predictors do not contribute to the results, and the analysis is conceptually equivalent to conducting separate regression equations for each group (level-2 unit), then analyzing these coefficients. For example, in a daily diary study in which daily Mood is being predicted by Stress, individual differences in mean daily stress would not influence the parameter estimates—in essence, they would be controlled. In contrast, if Stress were entered grand mean–centered, then estimates of the level-1 relationships between Mood and Stress would be influenced by level-2 differences in mean stress because daily mood would be modeled as a function of how much daily stress deviated from the grand mean of stress. The argument against group-mean centering of predictors is that this takes level-2 differences in predictors out of models, and does not represent the data properly. Regardless, the point remains that when level-1 predictors (either categorical or continuous) are grand mean–centered, the resulting intercepts are adjusted for level-2 differences in predictors. Whether such an adjustment is desirable will vary as a function of the questions of interest and the data at hand.

Comparing Fixed Effects

This section introduces what I think is one of the most powerful but underutilized aspects of MLM—the ability to compare coefficients through the use of constraints on a model.

Through the judicious use of coding, centering, and adding a nested level representing a measurement model, analysts can specify models in which coefficients represent very specific entities (means or relationships), and these entities can in turn, be compared, and very precise conclusions can be drawn.

A constraint on a model consists of applying weights to a group of coefficients (i.e., constraining them in some way), and the impact of this constraint on the fit of the model to the data is examined. If constraining the coefficients as represented by the weights results in a poorer fit of the model, then the comparison represented by the weights is assumed to be significant. The value of the constraint is the sum of the products of each coefficient and the weight assigned in the contrast to each coefficient. Note that the significance test is always a test against 0, specifically, a χ^2 test.

Assume a study of daily mood and daily stress, and the question of interest concerns differences in the strength of the mood–job stress and mood–family stress relationships. A model of such analyses is below:

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}\,(\text{Job}) + \beta_{2j}\,(\text{Family}) + r_{ij} \\ \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{2j} \end{aligned}$$

The mood–job stress and mood–family stress relationships are represented by the γ_{10} and γ_{20} coefficients, respectively, and the strength of these two coefficients can be tested by constraining them to be equal. In terms of weights, a simple weighting of 1 and –1 (in this case, it does not matter to which coefficient the weights are applied) tests the hypothesis that the two coefficients are the same. For example, if both slopes were –.25, then the constraint would sum to 0: $1^*(-.25) + -1^*(-25) = 0$, and the constraint result would not change the model fit. In contrast, if one of the slopes was –.50 and the other was –.25, then the constraint would not sum to 0, and it might be significantly different from 0.

Constraints can involve any number of coefficients. For example, if there were three level-1 predictors, the first two could be compared to the third with weights of 1, 1, and –2. In addition, constraints can involve multiple comparisons simultaneously. With three predictors at level 1, a single constraint could consist of two comparisons, such as 1, –1, 0 and 1, 0, –1. Testing multiple comparisons simultaneously helps to control Type I errors, although similar to the results of an analysis of variance (ANOVA) with multiple groups, on the basis of constraints with multiple comparisons, it is not possible to know exactly which coefficients differ from each other. Note that the degree of freedom (*df*) of the χ^2 test of the constraint corresponds to the number of comparisons in the constraint. When there is one comparison, the test has 1 *df*; when there are two comparisons, it has 2 *df*, and so forth.

The previous examples describe ways of determining whether coefficients (or combinations of coefficients) differ from each other per se, which is likely to be the focus of most, but *not all*, questions. For example, a researcher may be interested in the relative strength of two coefficients, irrespective of their sign (i.e., a comparison of the absolute value). This could be the case in a study of relationships between a daily outcome, such as self-esteem, and two other daily measures, such as positive events and negative events, as in the level-1 model below:

$$y_{ij} = \beta_{0j} + \beta_{1j}\,(\text{Positive}) + \beta_{2j}\,(\text{Negative}) + r_{ij}$$

For sake of argument, assume that the coefficient for positive events is +.25, and the coefficient for negative is –.25. Using the "standard" weights of 1 and –1 described previously, these two coefficients might be significantly different from each other (1*.25 + –1*–.25 = .50). Nevertheless, it is quite clear that their absolute values are not different—they are identical. To compare the absolute values, weights of 1 and 1 can be used. Note that when the coefficients are of equal magnitude, +.25 and –.25, the sum of the products is 0 (1*.25 + 1*–.25 = 0). If the coefficients were of different magnitude (e.g., +.25 and –.50), the constraint might be significant because it would not be 0 (1*.25 + 1*–.50 = –.25). For an application of this technique see Nezlek and Plesko (2001).

Coding

For present purposes, I consider two types of codes, dummy- and contrast-codes. *Dummy-codes* occur when a measure is represented by a 1 (usually indicating the presence or existence of something) and a 0 (indicating the lack of something). *Contrast-codes* consist of some type of comparison, with the limitation that the values that make up the contrast sum to 0. To compare two entities, a contrast code of 1 and –1 could be used. For three entities, a code of 1, 1, and –2 could be used. The sum of the weights in a contrast-code need to sum to 0 because in MLM, significance tests test coefficients against 0. If the coefficient for a contrast is significantly different from 0, then the differences represented by that contrast are significant.

Contrast- and dummy-codes provide different advantages, and I discuss these advantages within the context of a social interaction diary study in which interactions are nested within persons. The hypotheses concern differences in how intimate people find interactions that involve and do not involve a close friend, and interactions are classified as a function of whether a close friend was present or not.

This could be done using a contrast variable at the interaction level (level 1), and modeling intimacy as a function of this contrast variable.

$$y_{ij} = \beta_{0j} + \beta_{1j}\,(\text{Close Friend – Contrast}) + r_{ij}$$

If the contrast variable was entered as uncentered, the resulting intercept would represent the intimacy for interactions that were "neutral," and the slope would represent the difference. In turn, these coefficients could be brought up to level 2 and modeled as a function of an individual-difference measure, such as Neuroticism.

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\,(\text{Neuroticism}) + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}\,(\text{Neuroticism}) + u_{1j}$$

What is critical about this model is that the level-1 slope represents the difference in intimacy between interactions with friends present and interactions with no friends present, and accordingly, when this slope is modeled at level 2, individual differences in a difference score are being analyzed. See Schaafsma, Nezlek, Krejtz, and Safron (2010) for an application of such a model in which the level-1 model compares the interactions of ethnic/minority members that involved or did not involve a member of their ethnic group.

Although estimating such a difference score can serve many purposes, it cannot be used to test some hypotheses. Continuing the present example, a researcher might be interested in knowing whether the relationship between Neuroticism and Intimacy with friends is the same as the relationship between Neuroticism and Intimacy with nonfriends. What is needed to test such a hypothesis are separate estimates of the intimacy people find with friends and without them, and a difference score does not provide such estimates.

Providing separate estimates for level 1 requires using a set of dummy-codes in what is sometimes referred to as a *no-intercept* or *zero-intercept model*. In such models, the intercept is dropped, and the outcome is modeled as a function of a series of dummy-codes (one representing each category of a classificatory system). To continue the example, a model such as that below could be run.

$$y_{ij} = \beta_{1j}\,(\mathit{Friend}) + \beta_{2j}\,(\mathit{NoFriend}) + r_{ij}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}\,(\text{Neuroticism}) + u_{1j}$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}\,(\text{Neuroticism}) + u_{2j}$$

In the level-1 model, *Friend* is a dummy-coded variable indicating whether a close friend was present, and *NoFriend* is the corresponding variable indicating whether no close friend was present. Understanding how such a model estimates means for these two types of interactions involves estimating expected values. For interactions in which a friend was present, *Friend* = 1, and *NoFriend* = 0, and the expected outcome is the β_{1j} coefficient, because the *NoFriend* coefficient drops out. In contrast, for interactions in which a friend was not present, *Friend* = 0, and *NoFriend* = 1, and the expected outcome is the β_{2j} coefficient because the *Friend* coefficient drops out. Determining whether the relationship between Neuroticism and Intimacy with friends is the same as the relationship between Neuroticism and Intimacy with nonfriends is done at level 2 by constraining the γ_{11} and γ_{21} coefficients to be equal.

This type of dummy-coded analysis can be used at any level of a model, with any number of categories. Each category needs to be represented with its own code, and two categorical systems can be combined to provide the basis for more complex comparisons. Assume a study in which participants are men and women who either received some type of treatment or not, and the dependent measure is daily mood. If sex and treatment are combined to create a four-category system, any combination of groups can be compared to any other using tests of fixed effects. The level-2 predictors are four dummy-codes, each representing a combination of sex and treatment (yes/no). The model is below.

$$y_{ij} = \beta_{0j} + r_{ij}$$
$$\beta_{0j} = \gamma_{10}\,(\text{Men–Yes}) + \gamma_{20}\,(\text{Men–No}) + \gamma_{30}\,(\text{Women–Yes}) + \gamma_{40}\,(\text{Women–No}) + u_{0j}$$

In summary, to conduct these types of models the following conditions need to met.

1. All observations need to be classified unambiguously into a single category. There can be no overlap or dual-classification, nor can observations not be represented.
2. All categories (dummy-codes) need to be included in the model, and the intercept needs to be dropped.

Interactions, Moderation, and Mediation

In discussing moderation and mediation, I rely on the classic distinction discussed by Baron and Kenny (1986). Although their analytic framework has been criticized (and various alternatives have been proposed), I am more concerned with the conceptual issues they raised. *Moderation* exists when the relationship between two variables varies as a function of a third variable, whereas *mediation* is said to occur when a third variable explains the relationship between two others. Within the multilevel context, moderation is typically examined by some sort of interaction, and cross-level interactions are sometimes referred to as *moderating relationships*.

Interactions (and, by extension, moderation) can also be examined within levels of analysis. For guidance about how to do this, consult Aiken and West (1991). The same techniques can be used to examine within-level interactions at any level of analysis, albeit with some differences reflecting the need to group-center variables at lower levels of analysis. An example of a within-level interaction is presented in the section below on conceptualizing the multilevel structure. The models described there can be used to determine whether a level-1 slope (e.g., a mood–stress relationship) varies as a function of (is moderated by) a categorical level-1 variable (e.g., weekday vs. weekend).

It is also possible to examine within-level interactions between two continuous variables, as in Nezlek and Plesko (2003). In this study, measures of self-esteem and positive and negative events were collected each day. The analyses examined a buffering effect—the possibility that relationships between self-esteem and negative events would be weaker on days when more positive events had occurred than on days when fewer positive events had occurred. The interaction term was created by centering each event score within each individual (subtracting the mean for each person from each of his or her observations), then multiplying these centered scores. The resulting interaction term was entered as uncentered (because the measures on which it was based had been centered), and the two event scores (positive and negative) were entered group mean–centered. The resulting interaction term was significant, and estimated values for self-esteem on days that were ± 1 *SD* on positive and negative events (based on the within-person *SD* for events) indicated that positive events did buffer the effect of negative events on self-esteem.

Evaluating mediation within the multilevel context (particularly at level 1) is much more complex, and techniques to do this are being developed. Regardless of the specific method or focus, I recommend caution when evaluating mediation on the basis of changes in variance estimates. As noted in the section on effect sizes below, error variances in the multilevel context are not interchangeable with residual variance estimates within the OLS context. Moreover, when considering how a series of level-2 measures might mediate each other as predictors of a level-1 coefficient, it must be kept in mind that variance estimates (no matter how flawed they might be) can be used only when a coefficient is modeled as randomly varying. For an informed discussion of mediation within the multilevel context, see Bauer, Preacher, and Gil (2006), Card (Chapter 26, this volume) and Eid, Courvoisier, and Lischetzke (Chapter 21, this volume) for discussions of evaluation of mediation using structural equation modeling (SEM).

At level 2 (or at the highest level of a model), I think it is safe to follow the guidelines originally offered by Baron and Kenny (1986). A variable is said to mediate the relationship between two others if it is related to the predictor, and if, when added to the model, the original predictor becomes nonsignificant and the mediator is significant. This begs the question of partial mediation, which relies upon changes in variance estimates, which may be problematic. See Bauer, Preacher, and Gil (2006) for a discussion.

The most complex situation is lower level mediation, in which a level-1 variable mediates the relationship between two other level-1 variables. The possibility of moderated mediation (i.e., that mediated relationships themselves might vary as a function of level-2 variables) also needs to be taken into account. The most thorough treatment of this I have encountered is that provided by Bauer and colleagues (2006). Space does not permit a detailed description of their strategy, but they devised a way to estimate the direct and indirect effects of a mediator on an outcome, taking into account the possibility that these effects might vary across units of analysis. Their technique is a bit complex, and it may take those who are not experienced modelers a bit of time to understand how to apply to their own data, but it appears at present to be the best available method.

Conceptualizing the Multilevel Structure: What Should Be Nested within What and When?

In most instances, the multilevel structure of a dataset can be determined straightforwardly (e.g., in a diary-style study, days or interactions might be treated as nested within persons). Nevertheless, there may be instances in which the structure is not clear, or when different ways of structuring the data seem reasonable. For example, should a study in which multiple observations are collected each day (e.g., a beeper study) be conceptualized as a three-level model (observations within days within persons) or as a two-level model (observations within persons, ignoring days)?

As discussed previously, within a multilevel data structure, a level of analysis represents a sample of some kind, and when considering whether a level of analysis should be included in a model, the extent to which a sample would be adequately represented by that level needs to be taken into account. For example, if multiple observations were collected each day for a sample of people, but only 2 days of data were collected for each person, nesting observations within days and days within persons would mean that there would be only two units of analysis at the day level (only 2 days for each person). Two days do not provide a basis for estimating the day-level variance, so in such a case, the day-level variance could not be modeled. Moreover, unless an analyst was willing to assign some meaning to the fact that 1 day of data collection was the first day and the other day was the second (treating the days as fixed effects), there would be no basis on which the data from different people could be organized together in terms of the day on which data were collected. In such a case, observations would simply be nested within persons.

Such an example raises the issue of how many units are needed to constitute a level of analysis. Unfortunately, to my knowledge, there are no clear or firm guidelines for this. Given this, I recommend thinking of each level of analysis as representing a sample, then thinking of how many observations are needed to provide a basis for making an inference to a population. Certainly two observations are too few, and 15 are probably adequate for most purposes. The exact number will depend upon the parameters of interest. Because means are more reliable than covariances (slopes), estimating means reliably will take fewer observations than estimating slopes reliably. Regardless, when there are not enough units to constitute a level of analysis, but there are enough that an analyst does not want to ignore the matter completely (e.g., 5 days in a situation such as the example just discussed), the level can be included in the model with the understanding that it will not be possible to estimate reliably the variance at that level of analysis.

A similar but slightly different situation arises when units of analysis can be distinguished or classified in terms of a fixed number. For example, assume a study in which observations are collected each day, and hypotheses concern differences between weekdays and weekends. Should such data be analyzed with a three-level model, days nested within type of day with persons? Not really. First, there is the fact that the "middle" level of analysis would have only two possible units (a day can occur only during the week or on a weekend) and, as discussed earlier, two does not provide a basis for drawing an inference to a population. But more important, the categories of weekday and weekend are not drawn from a universe of types of days—there are only two possibilities.

In such cases, type of day should be treated as a day-level variable, and differences between weekends and weekdays can be examined using two dummy-codes, one representing each type of day (see the previous section on coding). To compare day-level means the following model could be run. *Wday* is a dummy-coded variable indicating whether a day is weekday or not, and *Wend* is the corresponding variable for weekend days. Note that the level-1 model has no intercept, and predictors are entered uncentered. The difference between means for the two types of days is tested by constraining the γ_{10} and γ_{20} coefficients to be equal.

$$\begin{aligned} y_{ij} &= \beta_{1j}\,(Wday) + \beta_{2j}\,(Wend) + r_{ij} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{2j} \end{aligned}$$

Differences between weekends and weekdays in the relationships between a day level predictor and an outcome can be examined with an extension of this model. The first step in this analysis is to multiply *Stress* by the dummy-codes for weekdays and weekends. All four measures are entered into a zero-intercept level-1 model. The dummy-codes themselves are entered as uncentered, and the two stress measures are entered group mean–centered. It is also possible to group-center the stress measures before multiplying them, then enter the products uncentered. Regardless, the difference between the *Stress* slopes for the two types of days is tested by constraining the γ_{20} and γ_{40} coefficients to be equal.

$$\begin{aligned} y_{ij} &= \beta_{1j}\,(Wday) + \beta_{2j}\,(Wday^{*}Stress) + \beta_{3j}\,(Wend) + \beta_{4j}\,(Wend^{*}Stress) + r_{ij} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{2j} \\ \beta_{3j} &= \gamma_{30} + u_{3j} \\ \beta_{4j} &= \gamma_{40} + u_{4j} \end{aligned}$$

The same logic holds for other levels of analysis. Assume our typical study of days nested within persons, and hypotheses concern differences between people, such as sex differences or differences as a function of some experimental manipulation. It would not be appropriate to add a third level of analysis and nest persons within sex or within an experimental condition. Such variables are best represented as person-level variables, and sex differences and differences between conditions can be represented using dummy-codes in a fashion that is structurally similar to the level-1 example just presented.

Standardizing Measures

In comparison to most OLS analyses, MLM analyses provide only *unstandardized* estimates of coefficients. Nevertheless, many analysts (and reviewers and editors) want (or expect) a description of relationships that are standardized in some way. Although some have proposed ways of standardizing MLM coefficients using estimates provided in MLM analyses (e.g., dividing coefficients by variance estimates), my sense is that such techniques may not be justified statistically. At present, I recommend standardizing measures in advance and using the unstandardized coefficients provided by the analyses.

The most straightforward situation is standardizing measures at the highest level of analysis (e.g., level 2 in a two-level model). For many ambulatory assessment studies, this will be the person. So, if days were nested within persons (or if observations were nested within days that were nested within persons, as in a three-level model), individual differences such as personality or physiological measures could be standardized across persons. In turn, the unstandardized coefficients produced by the MLM analyses (at this level of analysis) would then be functionally standardized (i.e., a 1 unit change in a coefficient would represent the change associated with a 1 *SD* change). Moreover, standardizing measures in this way makes it easier to interpret comparisons of the strength of coefficients, as discussed in the section on comparing fixed effects. Of course, if such measures were not distributed normally, they might need to be transformed and then standardized. Analysts need to make such decisions based upon the measures at hand and the norms within their disciplines.

Decisions about standardizing at lower levels of analysis (e.g., level 1 in a two-level model or levels 1 and 2 in a three-level model) are a bit more complicated. One option is to standardize across the entire population of observations. For example, in a diary study in which observations were nested within days, a day-level measure such as daily mood could be standardized in terms of all days, ignoring the nested structure of the data. Moreover, standardizing all of the day-level measures in such a study would have the advantage of equating the total variance of different measures, which may make the interpretation of coefficients easier.

Nevertheless, it is *critical* to keep in mind that such population standardization at level 1 does not produce standardized coefficients in the same sense that standardizing at level 2 does; that is, level-1 slopes from such data *do not* represent a change in a predictor associated with a 1 *SD* increase in level-1 predictors. This is because standardizing in this way sets the total variance (the sum of the between and within variances) to 1.0, and it does not make all the level-1 variances the same for all level-2 units. In a diary study, it would not set to 1.0 the within-person standard deviation for each participant. Within-person standard deviations would still vary. It is possible to standardize measures within each level-2 unit of analysis (e.g., each person), but such standardization is not recommended, because it may eliminate important aspects of the data from the model.

For analysts who want to express their results in terms of some sort of standard metric, I recommend generating predicted values for units of observation ± 1 *SD* from the mean. For level-1 measures, the within-person standard deviation can be estimated by taking the square root of the level-1 variance from an unconditional model (whether the measure has been population standardized or not), and then multiplying scores representing ± 1 *SD* by the coefficients corresponding to each measure.

Understanding Effect Sizes within the Multilevel Context

For analysts accustomed to estimating effect sizes within the OLS context by relying on variance estimates, it may be difficult to understand the difficulties inherent in using variances to estimate effect sizes within the multilevel context. Moreover, different researchers have approached the subject differently, adding to the confusion. My recommendation is conservative and reflects the advice of Kreft and de Leeuw (two well-respected modelers), who noted, "In general, we suggest not setting too much store by the calculation of R_B^2 [level-2 variance] or R_W^2 [level-1 variance]" (1998, p. 119).

Within the multilevel context, the standard method (to the extent that a standard exists) of estimating effect sizes is similar to that used within the OLS framework. The residual variances from models with different predictors are compared, and the reduction in variance from a model with added predictors is subtracted from the residual variance of the "base" (or previous model), this difference is then divided by the variance in the base model. The resulting quotient is the percent of residual variance that has been "explained" by the added predictors. Moreover, such a procedure can be applied to both level-1 and level-2 variance estimates.

At first glance, this sounds all well and good—within the multilevel context one follows the same procedures as in the single-level (OLS) context. Unfortunately, in terms of simplicity, there are some important caveats that need to be discussed. First, keep in mind that for MLM, it is possible to add a significant level-1 predictor to a model that does not decrease the level-1 residual variance—something that cannot happen in OLS analyses. In contrast to OLS procedures in which the significance tests of coefficients are based on variance estimates, in MLM, variances are estimated separately from tests of fixed effects. Such a possibility calls into question the accuracy of using variance reductions as a means of estimating effect sizes. Second, it is not possible to use variances to estimate effect sizes when predicting fixed effects. For example, if a slope is not modeled as random, then there is no residual variance for that slope, making it impossible to calculate a reduction in variance for that slope. Third, when estimating effect sizes at level 1, predictors should be entered as *group-mean*-centered. If predictors are not group-mean-centered, then the level-1 variance estimates will contain level-2 variance of those predictors, undermining the validity of relying upon a reduction in level-1 variance (Kreft & de Leeuw, 1998, p. 119).

At this point in time, my sense is that this issue is far from resolved, and I urge analysts to be cautious when using reductions in variances to estimate effect sizes. It seems that more caution is necessary at level 1 than at level 2. More specifically, using reductions in variances to estimate effect sizes may be unreliable when level-1 predictors are not entered group mean–centered and when there are multiple level-1 predictors. At times, editors and reviewers may (blindly) insist on including effect sizes based upon variance estimates, irrespective of the possible problems inherent in such a procedure. In such cases, authors have little choice but to provide such estimates, although they may want to discuss in passing the possible problems inherent in such a procedure.

Depending upon the circumstances, analysts may be able to describe the strength of the relationships they have found using other procedures. For example, the magnitude of coefficients can be evaluated in terms of the standard deviation of the outcome: How many standard deviations does a slope represent? Such a criterion can be applied to either level-1 or level-2 slopes, although when predicting level-1 slopes, it will be difficult to estimate the standard deviation of a slope that is not modeled as randomly varying. Regardless, it is important to keep in mind that the variance estimates produced in MLM analyses are not the same as those produced in OLS regression style analyses.

Estimating Reliability of Scales Administered on a Repeated Basis

A defining characteristic of ambulatory assessment studies is the fact that measures are collected on repeated occasions from the same person, and such measures may consist of multi-item scales (e.g., a three-item measure of anxiety is administered every day). When this is the case, researchers may want to know how reliable a scale is; that is, how consistently does a set of items measure the same construct? Note that this does not address the issue of validity, which concerns the correspondence between the construct being measured and the construct that is intended to be measured; rather, it concerns the multilevel equivalent of a Cronbach's alpha.

When thinking of estimating reliability within the multilevel framework, I think it helps to think that the reliability of a set of items is like an estimate of the average correlation between pairs of those items. The reliability of a scale for which the items are highly correlated with each other will tend to be higher than the reliability of a scale for which the items are not so highly correlated with each other. Let's consider the simple example of a study in which a scale is administered once a day for multiple days.

With this in mind, it may be helpful to start with descriptions of how such reliabilities should *not* be estimated. First, the reliability of the intercept from an unconditional model of mean scale scores is *not* the item-level reliability of the items making up the scale. Within MLM, the reliability of an intercept indicates how consistent responses are within level-2 units—in this case, how consistent the scale scores (the means of the items) are within an individual. By definition and design, scale means ignore the within-occasion variance of items and ignore the multilevel equivalent of within-person correlations among the items. Consequently, the reliability of the intercept provides no information about the day-level consistency of responses to individual items. Second, it is wrong to calculate within-person means for a set of items, then calculate a Cronbach's alpha using these mean scores at the person level. Such a procedure confounds within- and between-person variance, and provides an estimate of nothing. Finally, it is wrong to calculate a Cronbach's alpha for each day of a study, then combine these alphas somehow. This assumes that Day 1 for Person 1 should be matched with Day 1 for Person 2, and so forth. Keeping in mind the idea that reliability conceptually resembles a correlation, such a procedure would be similar to calculating a correlation between two variables for each day of a study, then averaging those correlations to estimate the relationship between the two variables. Moreover, what happens when participants have different numbers of days, or when days cannot be matched?

The proper way to estimate the item-level reliability of a set of items within the multilevel framework is to add to the model what is sometimes referred to as a *measurement model.* For the typical diary study, this is done by nesting items within occasions of measurement, then nesting occasions within persons. The model equations for such an analysis are below, and an example of this procedure can be found in Nezlek and Gable (2001).

$$\begin{array}{lll}\text{Item level} & \text{(Level 1)} & y_{ijk} = \pi_{0jk} + e_{ijk} \\ \text{Day level} & \text{(Level 2)} & \pi_{0jk} = \beta_{00k} + r_{0jk} \\ \text{Person level} & \text{(Level 3)} & \beta_{00k} = \gamma_{000} + u_{00k}\end{array}$$

In this analysis, *i* items are nested within *j* days, which are nested within *k* persons. The item-level reliability of the scale is the reliability of the level intercept (π). This is an indication of how consistent the responses are within days (and within persons)—the multilevel equivalent of a Cronbach's alpha at the item level. An example of the level-1 data structure for such an analysis is presented below. In this example, data are presented for two persons (*A* and *B*), for 2 days (1 and 2), for a three-item scale. The column labeled "Resp" contains the response for a specific item.

Person	Day	Item	Resp
A	1	1	3
A	1	2	4
A	1	3	4
A	2	1	4
A	2	2	3
A	2	3	5
B	1	1	2
B	1	2	1
B	1	3	3
B	2	1	2
B	2	2	2
B	2	3	1

There are a few caveats for such analyses. First, just as is required in a standard (single-level) reliability analysis, all items need to be scored "in the same direction." Items that are meant to be reverse-scored before computing scale scores need to be reversed before estimating the reliability. Second, when a study has multiple scales, I recommend estimating the reliability of each scale separately. When multiple scales are analyzed together using this procedure, the reliability of each scale is influenced by the reliability and number items of the other scales in the analysis. A more detailed explanation of this is beyond the scope of this chapter, but it can be found in Nezlek (2010), and another approach to estimating reliability (based on SEM) can be found in Shrout and Lane (Chapter 17, this volume). The approach Shrout and Lane advocate assumes that occasions of measurement are fixed (rather than random), an assumption that may hold under certain circumstances.

Analyzing Multiple Outcomes Simultaneously

With a simple extension, adding a measurement as a level-1 dataset (e.g., nesting items within occasions and occasions within persons) can be used to analyze multiple outcomes simultaneously. One of the important advantages provided by such an analysis is the ability to compare the strength of the relationships different outcomes have with the same predictor. Assume a daily study in which three measures are collected each day—stress, depression, and anxiety—and anxiety and depression are measured with two items each. Note that to do this type of analysis, outcomes need to be measured with more than one item (more than one indicator in formal terms).

Further assume that hypotheses of interest concern differences in mean levels of anxiety and depression, and differences in the strength of the relationship between stress and depression and the relationship between stress and depression. One way to test these hypotheses would be to run separate models for anxiety and depression, and compare (somehow) the resulting coefficients. The critical word in this sentence is *somehow*. It is very difficult (and requires making assumptions that cannot be easily defended) to test the strength of coefficients generated by separate models.

In contrast, if a measurement level is added as level 1 to a two-level model (items nested within occasions, occasions nested within persons—structurally the same as the earlier reliability analyses), the comparisons mentioned above are straightforward. Let's assume anxiety and depression are measured with two items each. The critical features of such an analysis are that at level 1, a dummy-coded variable is added, representing each outcome (Anx and Dep), and the intercept is dropped from the level-1 model. Such a level-1 model "brings up" to level 2 an estimate of the mean score for the items for each outcome. The model equations for such an analysis are below:

Item level		$y_{ijk} = \pi_{1jk}\,(\text{Anx}) + \pi_{2jk}\,(\text{Dep}) + e_{ijk}$
Day level	(Anxiety)	$\pi_{1jk} = \beta_{10k} + r_{1jk}$
	(Depression)	$\pi_{2jk} = \beta_{20k} + r_{2jk}$
Person level	(Anxiety)	$\beta_{10k} = \gamma_{100} + u_{10k}$
	(Depression)	$\beta_{20k} = \gamma_{200} + u_{20k}$

Similar to the reliability analysis described earlier, in this analysis, there are *i* observations nested within *j* days, which are nested within *k* persons. As presented in the sample dataset below, for the measurement level (level 1) of the analysis, there are a total of four observations for each day, one observation for each of the items of the two measures. In addition to the data identifying each observation, for each level-1 observation there are three measures: a dummy-code indicating whether the item is an anxiety item (Anx), a dummy-code indicating whether the item is a depression item (Dep), and the response itself (Resp).

When Anx and Dep are entered uncentered as predictors of Resp, the π_{1jk} and π_{2jk} coefficients become estimates of the means for the three anxiety and three depression items, respectively. Mean levels of anxiety and depression can be compared by constraining the γ_{100} and γ_{200} coefficients to be equal.

Person	Day	Anx	Dep	Resp
A	1	1	0	4
A	1	1	0	5
A	1	0	1	2
A	1	0	1	3
A	2	1	0	5
A	2	1	0	7
A	2	0	1	3
A	2	0	1	6
B	1	1	0	4
B	1	1	0	5
B	1	0	1	2
B	1	0	1	3
B	2	1	0	5
B	2	1	0	7
B	2	0	1	3
B	2	0	1	6

This model can be expanded by adding predictors at different levels. For example, daily stress could be added at the day level (which is now level 2). The model equations for such an analysis are below:

$$\begin{array}{lll}
\text{Item level} & & y_{ijk} = \pi_{1jk}\,(\text{Anx}) + \pi_{2jk}\,(\text{Dep}) + e_{ijk} \\
\text{Day level} & (\text{Anxiety}) & \pi_{1jk} = \beta_{10k} + \beta_{11k}\,(\text{Stress}) + r_{1jk} \\
 & (\text{Depression}) & \pi_{2jk} = \beta_{20k} + \beta_{21k}\,(\text{Stress}) + r_{2jk} \\
\text{Person level} & (\text{Anxiety}) & \beta_{10k} = \gamma_{100} + u_{10k} \\
 & (\text{Anx-Slope}) & \beta_{11k} = \gamma_{110} + u_{11k} \\
 & (\text{Depression}) & \beta_{20k} = \gamma_{200} + u_{20k} \\
 & (\text{Dep-Slope}) & \beta_{21k} = \gamma_{210} + u_{21k}
\end{array}$$

The strength of the relationship between stress and anxiety can be compared to the strength of the relationship between stress and depression by constraining the γ_{110} and γ_{210} coefficients to be equal. Predictors can also be added to the person-level model to model individual differences in day-level relationships, and using constraints similar to those used in the previous examples, the strength of these person level relationships can be compared.

Nonlinear Outcomes

The previous discussion has concerned linear (continuous) outcomes, but it is not uncommon for diary-style research to concern nonlinear outcomes such as categorical responses. A simple example of this would be a dichotomous outcome: Did an individual experience a certain event during a day? More complicated examples include categorical outcomes with more than two categories, which might or might not be ordered (i.e., different categories might indicate more or less of an underlying variable), count data that are not normally distributed (how many times a day something occurred), and so forth.

Analyzing such outcomes follows the same logic and relies upon models similar to the analyses of linear outcomes, but there are important differences. These differences are due to the fact that nonlinear outcomes violate one of the most important assumptions underlying MLM analyses of linear outcomes—the independence of means and variances. For example, the variance of a binomial measure is Npq, where N is the number of observations, p is the probability of an occurrence, and q is equal to $1 - p$. This lack of independence means that if nonlinear outcomes are treated as linear, the resulting parameter estimates are likely to be inaccurate. Multilevel analyses of nonlinear outcomes are the multilevel equivalent of single-level logistical regression, and such analyses are sometimes referred to as *multilevel logistical regression* (MLR).

Detailed discussion of the available options to conduct MLR is beyond the scope of this chapter, but a few details merit attention. Most fundamentally, to eliminate the dependence of the variance on the mean, in MLR analyses, the level-1 model includes what is essentially a transformation function for the outcome. The specific function varies as a function of the nature of the nonlinearity of the outcome. For example, for a dichotomous outcome, a Bernoulli model with $n = 1$ can be used. The level-1 model for such an analysis is below.

$$\text{Prob}(y = 1|\beta_{0j}) = \phi$$

In this model, a coefficient, ϕ, 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 (person). How such analyses are described varies considerably across disciplines, and analysts will need to take into account the norms for their disciplines as they prepare papers.

A second issue that arises in some MLR analyses is the target of inference of the estimated coefficients. The critical issue is the extent to which an analyst is interested in estimating the population mean of a coefficient versus differences among level-2 units in coefficients. If estimating population averages is more of a focus (e.g., on what percentage of days people report being depressed), than what are called "population average" estimates are probably more appropriate. In contrast, if the focus is on individual differences in how often (percentage of days) people report being depressed, "unit-specific" estimates might be more appropriate. Blanket recommendations about which of these two will be appropriate for a given circumstance are not possible, and analysts are encouraged to consult published research in their home disciplines for guidance about selecting a set of parameters upon which they will rely.

Fortunately, building models (adding and centering predictors, making decisions about error structures, etc.) for MLR is very similar to building models for analyses of linear outcomes. Nevertheless, when interpreting the results of MLR analyses, it is important to keep in mind exactly what a coefficient represents, which can be complicated at times. For example, when the outcome is multinomial with k categories, there will be $k-1$ functions representing the log-odds of being in one category relative to being in what is called the "reference category." So, if participants can choose among yes, no, or maybe for a daily measure, then one function will represent the odds of responding "yes" relative to responding "maybe," and a second will represent the odds of responding "no" relative to responding "maybe." Furthermore, level-2 differences in these coefficients need to be interpreted with this in mind. For analysts who are not familiar with logistical regression, interpreting the results of MLR can be particularly challenging and time consuming. Another feature of MLR is that there is no level-1 variance estimate. Given the nature of the distributions of the outcomes, there cannot be.

Software Options

The increased interest in and use of multilevel analysis have been accompanied by an increase in the number of programs that can perform MLM, and such programs can be thought of in terms of two broad categories. There are "general-purpose" programs that can perform MLM and a wide variety of other analyses (e.g., Statistical Analysis Software [SAS], Statistical Package for the Social Sciences [SPSS]), and there are "single-purpose" programs designed to do only MLM (e.g., Hierarchical Linear Modeling [HLM]: Raudenbush, Bryk, Cheong, & Congdon, 2004; MLwiN: Rabash, Steele, Browne, & Goldstein, 2009). There is broad agreement (a functional consensus) regarding the computational algorithms underlying MLM analyses, so different programs provide the same results when the same models are specified because they use the same algorithms. Note, however, the phrase "when the same models are specified." Analysts who are not experienced with MLM, or not experienced with how to conduct MLM using a specific program, may have trouble specifying the model they wish to test.

For analysts who are not that familiar with MLM, I recommend using a single purpose program such as HLM. Setting up models and interpreting outputs tend to be more straightforward when using single-purpose programs than when using general-purpose programs. Because single-purpose programs were designed to do only MLM analyses, the interface and output have been designed to set up MLM models and display the results of MLM analyses directly and efficiently. In contrast, for general-purpose programs, setting up MLM models and displaying the results of MLM analyses are just two of many possible analyses, so user interfaces have not been designed specifically for MLM. Of the single-purpose programs available, I have found HLM to be particularly accessible in terms of setting up models and interpreting output. For example, in HLM, there is no need to create different centered versions of predictors. The centering is done by the program when the model is specified.

On the other hand, general-purpose programs offer at least two advantages over single-purpose programs. First, data preparation and analyses are all done with the same program. Single-purpose programs tend to have fewer options to transform data than do general-purpose programs (e.g., there are no data transformation options in HLM), and some analysts find it bothersome to transform data "outside" of a single-purpose program, then bring the data into that single-purpose program. Second, general-purpose programs may have more options for certain types of sophisticated analyses than do single-purpose programs. Such possibilities are particularly the case for SAS, in which analysts can combine PROC MIXED with other procedures and conduct advanced analyses, such as "Mixture Models," in which similarities among error structures are then used as a basis for categorizing observations (in our case, respondents). Moreover, general-purpose programs such as SAS tend to provide more options for modeling complex error structures, although it is important to note that MLwiN, a single-purpose program, can also model some fairly sophisticated error structures.

When considering software options, I recommend that analysts make certain they fully understand all the components of the output of their programs (e.g., where exactly are the results of the tests of the fixed and random effects?). Although such a recommendation may seem foolish, I have reviewed papers and heard talks in which authors and speakers have focused on or discussed aspects of their outputs that were not truly relevant to the substantive questions at hand. Moreover, I suspect that such problems are more likely with general-purpose program than with single-purpose programs because of the differences in how results of analyses are organized in different types of programs. Finally, by their nature, multilevel analyses tend to be more complex than single-level analyses, and I recommend that analysts proceed cautiously when they consider adding sophisticated options. Less is not only more in terms of the number of predictors in a model, but this is also the case in terms of the sophistication of the model in the general sense.

Guidelines for Publication

Research using intensive repeated measures designs is conducted by scholars in a wide variety of disciplines. Norms about how to present the results of analyses probably vary widely, and the following guidelines need to be considered in light of such differences. As concisely as possible, readers should be given information that will allow them to understand the data structure, the models used, and the results.

1. The structure of the data should be described unambiguously. What is nested within what? The numbers of observations at each level of analysis should be presented, and the distribution of observations at each level should be provided. If days are nested within persons, the mean number and standard deviation of the days (per person) included in the analyses should be presented. If observations are also nested within days, the mean number and standard deviation of the number of observations per day should be presented. Criteria for including or excluding cases (persons, days, observations, etc.) should be described explicitly, and the number of excluded cases should be presented.

2. Multilevel summary statistics (means and variance estimates for each level of analysis) should be provided for all measures. Such statistics provide a context for understanding results. Similarly, univariate summary statistics should be presented for measures that were not nested.

3. When presenting results, in addition to the *p*-value, because $t = \gamma/SE$, it suffices to present two of the following: γ (or β), the *t*-value, and the standard error. I present γ and *t*.

4. At present, I recommend presenting the equations representing the models to clarify what predictors were included at each level of analysis. Moreover, à la Bryk and Raudenbush, I recommend presenting the equations for each level of analysis separately. Such separation emphasizes differences in phenomena across levels of analysis and highlights cross-level effects, something that is particularly helpful for readers who are not familiar with multilevel modeling.

5. How predictors were centered should be described explicitly (including uncentered). Results (coefficients) cannot be understood without knowing how predictors were centered. Do not assume that readers will know how predictors were centered.

6. The error structure of the model should be described, including a description of the basis used to fix effects that were not modeled as random. Furthermore, unless hypotheses explicitly concern error structures per se, detailed discussions of error structures are probably not necessary.

7. I strongly recommend that authors present predicted values when describing results. This can be very helpful to readers when models are complex, such as when there are cross-level interactions. For continuous measures, the standard is to generate predicted values for observations ± 1 *SD*. For categorical measures, predicted values can be generated for observations that fall into different categories.

8. Although it is a tradition in some disciplines, I do not necessarily see the value in presenting sequential tests of model fits. The emphasis in MLM (compared to SEM) is on the fixed effects rather than on the fit of the model as a whole. Moreover, most hypotheses concern fixed effects, and indices of model fits include both fixed and random components. When sequential comparisons provide insights, they are certainly valuable. When they do not, they tend to distract more than they inform.

Concluding Thoughts

A chapter such as this cannot cover in detail all the topics that are relevant to using MLM to analyze diary-style data. For one, I did not discuss different options for modeling error structures. Moreover, the importance of modeling more sophisticated error structures than the standard MLM structure (modeling each random error term and all the covariances among them) will vary across disciplines. For example, autoregressive structures seem to be more important for growth curve analyses than for data that are simply collected across time (e.g., in a daily diary study). There are numerous sources for information about modeling more complex error structures. A good place to start is Hox (2002).

I also did not discuss lagged analyses. Examining relationships between lagged coefficients (e.g., a measure at time *n* and another at time *n* + 1) can provide some insights into causal relationships between measures. I know of no formal treatment of lagged relationships within the multilevel framework; however, interested readers can consult Nezlek (2002) for an example of lagged analyses, with the understanding that the approach taken there is preliminary. Also of note is how I discussed mediation. The topic of mediation per se is a "hot" topic, with all sorts of back and forth on the hows and whys, and mediation within the multilevel context is part of this dynamic area of research. The Bauer and colleagues (2006) reference represents current good, perhaps best, practice, although my sense is that the issue is not fully resolved.

Regardless, this chapter should serve as a useful introduction to using MLM to analyze the data collected in diary-style studies for those unfamiliar with such applications. Persons truly and totally unfamiliar with MLM may want to complement their reading by consulting some of the references I mentioned at the beginning of the chapter. For those more familiar with MLM, some of the techniques and issues I have discussed may cause them to reevaluate their present practice. Such reevaluation may not lead to changes in practice, but it may lead to more thoughtful application of these techniques.

References

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

Baron, R. A., & Kenny, D. M. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations *Journal of Personality and Social Psychology, 51*, 1173–1182.

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.

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

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.

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

Kleinbaum, D. G., & Klein, M. (2002). *Logistic regression: A self-learning text* (2nd ed.). New York: Springer-Verlag.

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

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.

Nezlek, J. B. (2002). Day-to-day relationships between self-awareness, daily events, and anxiety. *Journal of Personality, 70*, 249–275.

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

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

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

Nezlek, J. B. (2010). *Understanding reliability in multilevel models.* Unpublished manuscript, College of William and Mary, Williamsburg, VA.

Nezlek, J. B. (2011). Multilevel modeling for social and personality psychology. In J. B. Nezlek (Ed.), *The Sage Library in social and personality psychology methods.* London: Sage.

Nezlek, J. B., Kafetsios, K., & Smith, C. V. (2008). Emotions in everyday social encounters: Correspondence between culture and self-construal. *Journal of Cross-Cultural Psychology, 39*, 366–372.

Nezlek, J. B., & Plesko, R. M. (2001). Day-to-day relationships among self-concept clarity, self-esteem, daily events, and mood. *Personality and Social Psychology Bulletin, 27*, 201–211.

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.

Rabash, J., Steele, F., Browne, W. J., & Goldstein, H. (2009*). A user's guide to MLwiN, v2.10.* Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

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

Schaafsma, J., Nezlek, J. B., Krejtz, I., & Safron, M. (2010). Ethnocultural identification and naturally occurring interethnic social interactions: Muslim minorities in Europe. *European Journal of Social Psychology, 40*, 1010–1028.

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

Wheeler, L., & Reis, H. (1991). Self-recording of everyday life events: Origins, types, and uses. *Journal of Personality, 59*, 339–354.