Multilevel Random Coefficient Analyses of Event- and Interval-Contingent Data in Social and Personality Psychology Research

John B. Nezlek
College of William & Mary

Increasingly, social and personality psychologists are conducting studies in which data are collected simultaneously at multiple levels, with hypotheses concerning effects that involve multiple levels of analysis. In studies of naturally occurring social interaction, data describing people and their social interactions are collected simultaneously. This article discuses how to analyze such data using random coefficient modeling. Analyzing data describing day-to-day social interaction is used to illustrate the analysis of event-contingent data (when specific events trigger or organize data collection), and analyzing data describing reactions to daily events is used to illustrate the analysis of interval-contingent data (when data are collected at intervals). Different analytic strategies are presented, the shortcomings of ordinary least squares analyses are described, and the use of multilevel random coefficient modeling is discussed in detail. Different modeling techniques, the specifics of formulating and testing hypotheses, and the differences between fixed and random effects are also considered.

The use of nonexperimental methods in social and personality psychology has increased dramatically over the past two decades, and much of this research has concerned day-to-day experience and other naturally occurring phenomena. Many studies of such phenomena produce multilevel data structures in which observations are collected simultaneously at multiple levels, and this article describes the use of multilevel random coefficient modeling to analyze data generated in studies of naturally occurring phenomena. An important way to classify studies of naturally occurring phenomena is the basis used to structure the data, and two types of data structures, event and interval contingent, predominate the study of naturally occurring phenomena in social and personality psychology. In event-contingent data structures, a specific type of event triggers (or organizes) data collection, whereas in interval-contingent structures, data are collected at intervals, such as every day.

A wide variety of events and various intervals can be used to organize data collection, and it is beyond the scope of this article to discuss the analysis of each of these different data structures. In light of this, the analysis of event-contingent data is discussed in terms of the study of day-to-day social interaction, and the analysis of interval-contingent data is discussed in terms of the study of reactions to daily events. These topics were chosen because they are popular topics among social and personality psychologists that are frequently studied with multilevel designs. In studies of day-to-day social interaction, each person describes the social interactions (i.e., events) that occur over a specified period of time, and interactions are treated as nested within people. In studies of daily events, people provide data at a certain interval (usually each day), and measures for days (an interval) are treated as nested within people. In the terminology of multilevel analysis, interaction- and day-level analyses are micro-level analyses, and person-level analyses are macro-level analyses.

Although this article focuses on the analysis of event- and interval-contingent data, the principles and techniques discussed in this article are relevant to the analysis of various other types of data structures. Multilevel analyses are appropriate for any type of hierarchically struc-

Author's Note: Preparation of this article was supported by a grant from the College of William & Mary faculty research program. I am grateful to William Cunningham, Lee Kirkpatrick, Mathilda duToit, and Linda Zyzniewski for their comments on previous versions of this article. Correspondence regarding this article should be sent to John B. Nezlek, Department of Psychology, P.O. Box 8795, College of William & Mary, Williamsburg, VA 23187-8795, e-mail: john.nezlek@wm.edu.

PSPB, Vol. 27 No. 7, July 2001 771-785
© 2001 by the Society for Personality and Social Psychology, Inc.

tured data in which there is some dependency (or lack of independence) among observations. For example, a researcher might collect data in groups, and analyses of such data should account for the variation due to the groups within which observations were collected (e.g., Nezlek & Zyzniewski, 1998). Similarly, observations might be nested within couples, and analyses of such data should account for the variation due to the couples within which observations were collected (e.g., Barnett, Marshall, Raudenbush, & Brennan, 1993).

This article was prepared primarily for researchers who have some experience with multilevel designs but are relatively unfamiliar with recent advances in multilevel analysis. No familiarity with these newly developed techniques was assumed, although interested readers will probably benefit by consulting a more detailed treatment of multilevel analytic techniques to complement the material presented in this article. Three excellent introductions to multilevel analysis are Bryk and Raudenbush's (1992) *Hierarchical Linear Models*, Kreft and de Leeuw's (1998) *Introducing Multilevel Modeling*, and Snijders and Bosker's (1999) *Multilevel Analysis*. In addition, applications of multilevel analyses to the study of personal relationships are discussed by Gable and Reis (1999), and Affleck, Zautra, Tennen, and Armeli (1999) discuss the application of multilevel analyses to daily process studies (i.e., daily event studies in the present context).

ANALYTIC STRATEGIES FOR MULTILEVEL DATA STRUCTURES

The sine qua non of multilevel analysis is that phenomena are examined at different levels of analysis simultaneously. For example, assume that two state variables, self-esteem (SE) and feelings of belongingness (BL), are measured multiple times for a sample. Within such a data set, the relationship between SE and BL can be examined at both the within- and between-subject levels. At the within-subject level, the question is whether, for any person, BL is higher when SE is higher, and at the between-subject level, the question is whether the BL scores of people are higher if their SE scores are higher.

When examining such relationships, it is critical to recognize that relationships at the between- and within-subject levels are independent. As illustrated in Figure 1, within-subject relationships can be negative, whereas between-subject relationships are positive; within-subject relationships can be positive, whereas between-subject relationships are negative; or there can be no relationship at the within-subject level, whereas there is a relationship at the between-subject level. Moreover, these three examples do not exhaust the possibilities. Any type of relationship at one level can coexist with any type of relationship at another level, and techniques that do not examine phenomena at different levels simultaneously can provide misleading descriptions of the relationships within a data set.

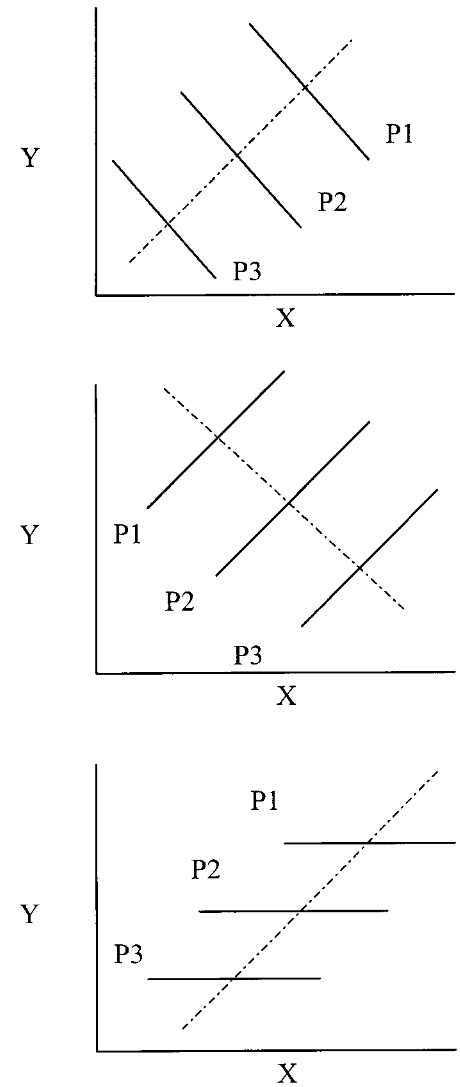

Figure 1
NOTE: Solid lines represent within-subject relationships for three different people: P1, P2, and P3. Dashed lines represent between-subject relationships (relationships between mean X and Y).

For various reasons, it is not appropriate to use traditional ordinary least squares (OLS) procedures to analyze the multilevel data collected in studies of naturally occurring phenomena. For example, unless all participants have the same number of observations, repeated measures analyses of variance (ANOVAs) cannot be used. Moreover, as discussed below, even if participants have equal numbers of observations, OLS ANOVAs can produce inaccurate estimates of error terms and have other shortcomings. Nevertheless, multilevel data structures have been analyzed using OLS procedures, and such analyses generally fall into one of two categories—disaggregation and aggregation procedures—and it is instructive to review briefly these techniques and their shortcomings.

DISAGGREGATION AND AGGREGATION ANALYSES

In disaggregation analyses, characteristics of macro-level units of analysis (people in social interaction and daily event studies) are assigned to micro-level units of analyses (interactions or days), and analyses are done only at the micro level. For example, in a study of daily social support by Cutrona (1986), participants described the supportive interactions they had each day and provided other daily measures. The data were analyzed using an OLS regression technique in which the day was the unit of analysis, and the social support that occurred each day comprised the dependent variables. Differences in participants' mean ratings were partialed out through the use of n-1 dummy-coded independent variables, and relationships were examined by assigning subject-level variables to each of a person's interactions. One of the major shortcomings of such approaches is that they do not examine how relationships between variables may vary across people. For example, in Cutrona, the analyses could not examine how the relationship between daily depressed mood and receipt of social support varied across individuals.

Aggregation analyses in which summary measures are calculated for each macro-level unit of analysis (e.g., person) and analyses are done at the person level provide some advantages over disaggregation analyses. One advantage is that aggregation analyses can examine how relationships at one level of analysis vary across another level. For example, Wheeler and Nezlek (1977) examined how a within-person phenomenon (differences in reactions to same-, opposite-, and mixed-sex interaction) varied as a function of participant sex. They did this by calculating separate scores for each person describing their same-, opposite-, and mixed-sex interactions and then analyzed these means using repeated measures OLS ANOVAs with type of interaction as within-subject factor and participant sex as a between factor. Moreover, other types of person-level aggregation have been used in studies of social interaction. For example, Hodgins, Koestner, and Duncan (1996) calculated within-subject correlations among different ratings of interactions and then correlated these correlations with person-level variables. Studies of daily events have also relied on within-person summary measures such as within-person comparisons of mean affect on days when social events occurred with days when they did not (e.g., Clark & Watson, 1988) and within-person regression coefficients between mood and events (e.g., Affleck, Tennen, Urrows, & Higgins, 1994).

One problem with many aggregation analyses is that they do not take into account the reliability of the aggregated observations. For example, people tend to have more same- than opposite-sex interactions, and so means describing same-sex interactions may be more reliable than those describing opposite sex interactions because they are based on more observations. Such problems are even more important when estimating relationships between variables such as within-subject covariances (correlations or regression coefficients) because the reliability of a measure of covariance reflects the reliability (or lack thereof) of two measures. One popular solution to this problem has been the use of weighted least squares (WLS) analyses in which aggregated observations are weighted, for example, by their size and the precision of the estimated coefficient (Kenny, Kashy, & Bolger, 1998). Nonetheless, even such WLS analyses can provide misleading parameter estimates.

OLS ANALYSES AND RANDOM EFFECTS

The fundamental shortcoming of these analyses is that they do not conceptualize error properly. More specifically, they treat coefficients as fixed and not random and, in so doing, provide potentially misleading parameter estimates and tests of significance. For many analysts whose primary training and experience is analyzing data within the contexts of factorial experiments and single-level OLS regression, the distinction between fixed and random coefficients is not straightforward. Mathematically speaking, the distinction is relatively straightforward, however, and is a direct function of how the variance of a coefficient is modeled. A coefficient is referred to as fixed if a random error parameter is not modeled (estimated) for the coefficient, and it is referred to as random if one is.

Discussing fixed and random effects requires consideration of the inference space of an effect, the population a coefficient is meant to describe. Generally speaking, the broader the inference space, the more reasons to model a coefficient as random rather than fixed, which is an issue discussed by Littell, Milliken, Stroup, and Wolfinger (1996, pp. 230-234). For example, in most studies of day-to-day social interaction and daily events, little if any importance is placed on the specific interactions that are measured or on the exact days over which a study takes place. Interactions and days are (presumably) randomly sampled from populations of interactions and days and are meant to represent participants' typical lives. Presumably, coefficients based on samples of other days or interactions would be just as valid (although not exactly the same) as those based on the sample collected, and so coefficients are random in that they were sampled from each participant's population of possible coefficients. This sampling of coefficients constitutes a prima facie case for treating (modeling) coefficients describing within-person relationships such as interaction- and day-level phenomena as random, not fixed. Procedures that do not model such coefficients as random, such as the OLS and WLS procedures used in much research, may provide misleading parameter estimates because they do not model error properly.

MULTILEVEL RANDOM COEFFICIENT MODELING AND THE LOGIC OF MULTILEVEL ANALYSES

There is an emerging consensus that a technique known as multilevel random coefficient modeling (MRCM) provides the most accurate analyses of the types of multilevel data structures under consideration (e.g., Bryk & Raudenbush 1992; Kenny et al., 1998; Kreft & de Leeuw, 1998). One of the most important advantages of MRCM over comparable OLS procedures is its ability to model random error at all levels of analysis simultaneously, which is an advantage due to the fact that MRCM relies on maximum likelihood procedures to estimate coefficients.

Occasionally, multilevel analyses (with or without random coefficients) have been referred to as hierarchical linear modeling analyses. In keeping with the recommendation of de Leeuw and Kreft (1995), the label *MRCM* is used in this article to avoid the all too common confusion associated with referring to a type of analysis (MRCM) with the name of a widely used specific technique (HLM). de Leeuw and Kreft strongly recommend distinguishing HLM the model and program from hierarchical linear modeling the technique (or approach), and the label *MRCM* does this while highlighting the importance of estimating random (vs. fixed) coefficients.

In MRCM analyses, coefficients describing phenomena at one level of analysis are analyzed at another. In essence, a regression equation is estimated for each unit of analysis at one level, and these coefficients become the dependent variables in regression equations at the next level of analysis. Moreover, it is exactly how this is done that distinguishes different MRCM procedures from each other and from OLS multilevel analyses. In MRCM analyses, two parameters are estimated for each coefficient. The first, referred to as a fixed effect, is an estimate of the central tendency (mean) of a coefficient. The questions posed by most social and personality psychologists concern tests of fixed effects: On average, is a coefficient significantly different from 0? (Note that this does not examine the hypothesis that all coefficients are different from 0 or that all are less or greater than 0.) The second estimated parameter is the random error term associated with a coefficient, and it is also tested: Is the random error for a coefficient significantly different from 0? It is common in the modeling literature to discuss coefficients as random or fixed on the basis of whether the random error parameter is significant (a random effect) or not (a fixed effect). Deciding when to model a coefficient as random or fixed is discussed in more detail later.

Although multilevel analyses can have any number of levels, this article concerns two-level models. In standard MRCM nomenclature, Level 1 coefficients are referred to as βs (subscripted 0 for the intercept, 1 for the first coefficient, 2 for the second, etc.), and the basic Level 1 model is

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

In this model, there are *i* observations for *j* individuals of a continuous variable *y*, which are modeled as a function of the intercept for each person (β_{0j}, the mean of *y*) and error (r_{ij}), and the variance of r_{ij} is the Level 1 random variance. Such models are referred to as unconditional at Level 1 because *y* is not modeled as a function of another variable at Level 1. In a daily events study, *y* might represent a day-level measure such as daily affect, and in a social interaction diary study, *y* might represent an interaction-level measure such as intimacy of interaction.

Level 1 coefficients are then modeled (analyzed) at Level 2, and Level 2 coefficients are referred to as γs. There is a separate Level 2 equation for each Level 1 coefficient. The basic Level 2 (or macro-level) model is

$$\beta_{0j} = \gamma_{00} + u_{0j} .$$

In this model, the mean of *y* for each of *j* units of analysis (β_{0j}) is modeled as a function of only the grand mean (γ_{00}) and error (u_{0j}), and the variance of u_{0j} is the Level 2 variance. Such models are referred to as unconditional at Level 2 because β_{0j} is not modeled as a function of another variable at Level 2. In other words, no hypotheses (other than testing if the mean, γ_{00}, is significantly different from 0) are tested. In studies of daily events and social interaction, between-person relationships are typically modeled at Level 2.

Analysts are strongly advised to run a totally unconditional model before running other models. A totally unconditional model (also referred to as a null model) is one in which no term other than the intercept is included at any level. Although such models typically do not test hypotheses per se, they describe how much of the total variance of *y* is at each level. The total variance of *y* is the sum of the variances at each level. In a two-level model, this is the sum of the variance of r_{ij} and of u_{0j}, and the distribution of the total variance of *y* suggests the levels at which further analyses might be productive. For example, if all the variance in a measure is at Level 1, it may be difficult to model variance at Level 2.

ILLUSTRATIVE ANALYSES

These and other principles of MRCM will be illustrated by the analysis of a fictitious sample of data, contained in Appendices A and B. This data set describes 15 people (macro-level or Level 2 unit of analysis) who contribute up to 10 observations each, the micro-level or Level 1 units of analysis. Two variables, belongingness and state self-esteem, were measured at each observation (which could be interval or event contingent), and for each person, two variables were measured, adjustment and need for closure. In addition, a measure of adjustment multiplied by 10 is included to illustrate the importance of scale metrics. For present purposes, the observation-level (state) measure of self-esteem is the dependent measure. A totally unconditional analysis of state SE found that the Level 1 variance (the variance of r_{ij} from the Level 1 model) was 1.62, and the Level 2 variance (the variance of u_{0j} from the Level 2 model) was 3.00. The program HLM (Raudenbush, Bryk, Cheong, & Congdon, 2000) also provides reliability estimates for fixed effects, and the reliability of the intercept was .94. See Bryk and Raudenbush (1992, pp. 39-44) for a description of how HLM estimates reliability.

The next stage in most analyses is to examine conditional models. Conditional refers to the fact that an effect is modeled as a function of another variable. Such models are created by adding terms to equations at different levels. Using the data in the hypothetical data set, the following model analyzes y (state self-esteem) at Level 1 as a function of the other observation-level variable, belongingness (Belong):

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

For each of j individuals, a coefficient representing the relationship between self-esteem and belongingness is estimated (β_{1j}). In the multilevel modeling literature, such coefficients are referred to as slopes to distinguish them from intercepts. In a daily-events study, the negative events that occurred each day could be substituted for belongingness, in which case the slope would represent the relationship between y and negative events. In an interaction diary study, an interaction-level variable could be used to distinguish interactions that were dyads from those that were not.

The statistical significance of the relationship between self-esteem and belongingness (is the mean slope for belongingness across the j individuals different from 0?) is examined at Level 2 with an extension of the basic Level 2 model:

$$\text{Intercept: } \beta_{0j} = \gamma_{00} + u_{0j}$$

$$\text{Belong: } \beta_{1j} = \gamma_{10} + u_{1j}\,.$$

The mean slope is represented by γ_{10}, and if γ_{10} is different from 0, then the mean slope for belongingness is also. Note that both β_0 and β_1 are modeled as random effects; each has a random error term (u_{0j} and u_{1j}). To fix an effect, the random error term is deleted.

CENTERING OPTIONS

When adding predictors, analysts need to be mindful of the implications of different centering options. Centering refers to the location (or reference value) of a predictor used to estimate coefficients. For example, in standard OLS regression, variables are typically mean centered in that coefficients are based on deviations from sample means. In MRCM, three common options are no centering, grand mean centering, and group mean centering. Assume an analysis in which y is being modeled as a function of x at Level 1. If x is not centered, the intercept for each unit of analysis represents the expected score for y when x equals 0. If x is grand mean centered, the intercept is the expected score for y when x equals the grand mean. If x is group mean centered, the intercept is the expected score for y when x equals the mean of x for each unit of analysis. An important difference between grand and group mean centering is that with grand mean centering, intercepts are adjusted for differences in x across units of analysis.

Centering options are discussed in Bryk and Raudenbush (1992, pp. 25-29) and Kreft and de Leeuw (1998, pp. 106-114) and more formally in Kreft, de Leeuw, and Aiken (1995). Specific centering options in the analysis of interval and event-contingent data are discussed later, using the following rough guidelines. First, models with uncentered predictors are more likely than those with centered predictors to experience problems estimating parameters due to high correlations between intercepts and slopes, and frequently, predictors will need to be centered to reduce this correlation. Second, if 0 is not a sensible value for a Level 1 variable (e.g., the variable is a 1 to 7 scale on which 0 is not a possible value), then the variable should be centered. When a variable is grand mean centered, between-group differences in mean scores on the variable contribute to parameter estimates (including estimates of errors), whereas when the variable is group mean centered, they do not. Nevertheless, as Bryk and Raudenbush (1992) note, “No single rule covers all cases” (p. 27), and analysts need to make decisions about centering based on the structure of their data and the hypotheses of interest.

To illustrate the types of differences different centering options can create, the observation-level (Level 1)

TABLE 1: Parameter Estimates With Different Centering Options

Coefficient	*Parameter*	*No Centering*	*Grand Mean Centering*	*Group Mean Centering*
Intercept	Fixed	2.7554	5.0132	4.9785
	Random	9.4097	2.7567	3.1052
Reliability		0.909	0.967	0.974
Belong slope	Fixed	0.4017	0.4017	0.3942
	Random	0.0677	0.0679	0.0631
Reliability		0.722	0.723	0.704

relationship between self-esteem and belongingness in the hypothetical data set was examined with belongingness not centered, grand mean centered, and group mean centered. The results of these analyses are summarized in Table 1. Although centering influences all parameters in a model because of the interdependence of covariance estimates, for many of the analyses discussed in this article, centering will have the strongest influences on estimates of intercept parameters, both fixed and random effects.

The coefficients from these analyses are interpreted like the coefficients from OLS regression analyses. The mean slope (functionally equivalent to but not the same as the mean unstandardized regression coefficient between self-esteem and belongingness) was .4017 for the uncentered and grand mean analyses and .3942 for the group mean–centered analyses. For the uncentered and grand mean–centered analyses, an average (across persons) for every 1 point increase in belongingness, self-esteem scores increased .4017, and for the group mean–centered analyses the increase was .3942. Also, centering reduced the correlation between the intercept and the belongingness slope. This correlation was –.98 in the uncentered analysis, –.928 when belongingness was grand mean centered, and –.913 when belongingness was group mean centered.

Variables can also be added to Level 2 models to analyze Level 1 coefficients, either intercepts or slopes.[1] If Level 2 models describe person-level phenomena, then individual differences are analyzed at Level 2, and the guidelines for constructing variables follow those for OLS regression (e.g., Aiken & West, 1991). To examine Level 2 (individual) differences in Level 1 means, an unconditional Level 1 model is used so that the intercept represents the mean for each Level 1 unit (person):

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

The intercepts can then be modeled as a function of Level 2 variables of interest. In the test data set, adjustment (ADJ) was measured at Level 2, and the following model examines relationships between mean state self-esteem and adjustment:

TABLE 2: Analyses of Level 1 Coefficients

		Level 2 Coefficients	
		Intercept	*Adjustment*
Level 1 model: Intercept only			
Intercept	Fixed	4.982143	0.902788
	Random	1.38825	
	t ratio	15.409	3.924
	p level	.000	.002
Level 1 model: Intercept and slope			
Intercept	Fixed	4.981873	0.902489
	Random	1.48945	
	t ratio	15.392	3.922
	p level	.000	.002
Slope	Fixed	0.396176	-0.108294
	Random	0.04424	
	t ratio	5.736	2.203
	p level	.000	.046

$$_{0j} = \gamma_{00} + \gamma_{01}(\text{ADJ}) + u_{0j}\,.$$

Just as in OLS regression, the relationship between β_0 (the mean state self-esteem for each person, each Level 2 unit) and ADJ is tested by the significance of γ_{01} coefficient, and the results of such an analysis are summarized in Table 2. ADJ was significantly related to state self-esteem. For Level 2 variables, there are two centering options, none and grand mean. Grand mean is the option most comparable with standard OLS regression in that the intercept represents the predicted score for a unit of analysis at the mean for the independent variables (e.g., ADJ), and grand mean centering was used for this analysis.

Most MRCM programs estimate unstandardized coefficients, and this needs to be kept in mind when interpreting coefficients. In the present analysis, the intercept (4.982143) represents the state self-esteem score for a person who was at the mean on ADJ. The coefficient for ADJ (0.902788) means that for each unit increase (1 point) in ADJ, state self-esteem scores increase 0.902788. Note that centering options at Level 2, unlike those at Level 1, typically do not have important implications for convergence. Nevertheless, centering options are important for interpreting coefficients at Level 2. For example, when ADJ was not centered, the ADJ coefficient was the same as when it was grand mean centered; however, the intercept was –5.730944.

CROSS-LEVEL EFFECTS

The logic of analyzing intercepts from Level 1 models can also be extended to Level 1 slopes. How a Level 1 slope varies as a function of a Level 2 variable is frequently referred to as a either a cross-level interaction (because such terms concern how a relationship at one level varies as a function of variables at another level) or a slopes-as-outcomes analysis (because slopes from Level 1 become dependent measures at Level 2). Previously, it was found that self-esteem and belongingness covaried at Level 1.[2] To determine if this relationship (slope) varied as a function of a Level 2 variable (e.g., ADJ), the following model was analyzed, with Belong group mean centered:

$$\text{Level 1: } y_{ij} = {}_{0j} + \beta_{1j}(\text{Belong}) + r_{ij}.$$

At Level 2, these Level 1 coefficients are then modeled as a function of ADJ:

$$\text{Intercept: } {}_{0j} = \gamma_{00} + \gamma_{01}(\text{ADJ}) + u_{0j}\,.$$

$$\text{Belong: } \beta_{1j} = \gamma_{10} + \gamma_{11}(\text{ADJ}) + u_{1j}\,.$$

In the test data set, the esteem-belongingness slope was significantly related to ADJ $\gamma_{11} = -0.108294$), and so for every unit increase in ADJ, the slope decreased 0.108294 units. One way to describe this relationship would be to say that ADJ moderated the esteem-belongingness relationship.

COMPARING COEFFICIENTS

Frequently, researchers are interested in models in which more than one independent variable is included at a level of analysis, and MRCM analyses can easily accommodate more than one independent variable at any level of analysis. Moreover, coefficients in the same equations or across equations of a model can be compared. For example, in the test data set, the relative strength of relationships between state self-esteem and ADJ and between state self-esteem and need for control (NFC) can be examined by comparing the γ_{01} and γ_{02} coefficients in the following Level 2 model:

$$\text{Intercept: } \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{ADJ}) + \gamma_{02}(\text{NFC}) + u_{0j}\,.$$

In HLM, such comparisons are evaluated with χ^2 tests. An analysis of the test data set found that γ_{01} and γ_{02} coefficients (.956 and .022) were significantly different, $\chi^2(1) = 12.72$, $p < .001$. The background for such tests in HLM can be found in Bryk and Raudenbush (1992, pp. 48-56), and how to conduct such analyses is explained in Raudenbush et al. (2000). When making such comparisons, it is critical to keep in mind that the results of such tests can vary as a function of the metrics of the scales being used, which is a topic discussed in more detail later. Nevertheless, for the hypotheses examined by many personality and social psychologists, these tests are particularly appropriate because of their flexibility and ability to control for error rates inherent in multiple comparisons (individual contrasts can be tested as a group).

DETERMINING EFFECT SIZES

An important aspect of random coefficient modeling that is meaningfully different from OLS analyses is how the strength of an effect is determined. In OLS analyses, the same parameter (a sum of squares) serves as the basis for both significance tests and estimates of effect size, usually expressed as variance accounted for. In MRCM, tests of significance and variance estimates are based on different parameters. As discussed previously, significance tests are tests of fixed parameters, and in an important sense, a fixed-effect parameter represents the size of an effect (how much one measure changes as a function of another). In MRCM, estimating effect sizes in terms of variance accounted for relies on comparisons of variance parameters, using a model comparison procedure similar to that used in OLS (Bryk & Raudenbush, 1992, pp. 65-70).

To estimate the strength of the relationship between two variables at Level 1, the Level 1 error variances of two models are compared. (Note: These results are not presented in Tables 1 or 2 but can be obtained by analyzing the test data set.) The first model is unconditional, and the second includes a slope at Level 1. In the test data set, the Level 1 variance of state self-esteem from the unconditional model was 1.6813. When belongingness was included at Level 1 (group mean centered), the Level 1 variance was .740. This amounts to a reduction of 56%, corresponding to a correlation of .75. It is important to note that such a correlation is adjusted for differences in sample sizes, means, and reliability of measurement across units of analysis. To estimate the strength of the relationships between a Level 1 coefficient (intercept or slope) and a Level 2 variable, Level 2 error variances are compared. The random variance of a Level 1 coefficient from a model in which that coefficient is not modeled at Level 2 is compared with the random variance from a model in which it is modeled at Level 2. In the test data set, the variance of the intercept from the unconditional model was 3.004. When adjustment was included at Level 2, the variance for the intercept was 1.388. This amounts to a reduction of 54%, corresponding to a correlation of .73.

It is important to note that effect sizes can be estimated in this way only when effects are modeled as random. If no random error component is modeled, error variances cannot be compared. Moreover, as discussed by Kreft and de Leeuw (1998, p. 119), such calculations are better done with group mean–centered variables because such centering reduces problems with cross-level confounding of variances. Finally, analysts need to be aware that adding variables that have significant fixed effects does not necessarily increase the variance accounted for by a model (Kreft & de Leeuw, 1998). In sum, although Kreft and de Leeuw discuss R^2 in various contexts in their book, they advise analysts to be cautious when interpreting such estimates of effect sizes: "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]" (p.119).

INTERVAL-CONTINGENT DATA: REACTIVITY TO DAILY EVENTS

In a typical study of reactivity to daily events, participants describe the events that occur each day, and they provide measures of psychological states. Hypotheses of interest generally concern the day-level (or within-person) covariation between events and states and how this covariation varies as a function of trait-level individual differences such as neuroticism (e.g., Bolger & Zuckerman, 1995; Suls, Green, & Hillis, 1998). In a multilevel analysis, this means that slopes describing within-person relationships between daily states and daily events are estimated for every person, and individual differences in these slopes are then analyzed.

A basic within-person model is

$$y_{ij} = \beta_{0j} + \beta_{1j}\text{PosEvent} + \beta_{2j}\text{NegEvent} + r_{ij},$$

in which y is a daily score (e.g., mood) for person j on day i, β_{0j} is a coefficient representing the intercept for person j, β_{1j} PosEvent is a coefficient (a slope) for positive events, β_{2j} NegEvent is a coefficient (a slope) for negative events, and r_{ij} represents error. The influence on parameter estimates of between-person differences in numbers of events can be eliminated by group mean centering event scores.

Hypotheses about within-person relationships between states and events are examined by analyzing within-person slopes at the between-person level with a model such as

$$\text{Intercept: } \beta_{0j} = \gamma_{00} + u_{0j};$$

$$\text{Positive events: } \beta_{1j} = \gamma_{10} + u_{1j};$$

$$\text{Negative events: } \beta_{2j} = \gamma_{20} + u_{2j}.$$

In these models, γ_{00} represents the average of the day-level intercepts, and γ_{10} and γ_{20} represent the average of the positive and negative event slopes, respectively. Note that in this model, all three within-person coefficients are modeled as random (i.e., the u_{0j}, u_{1j}, and u_{3j} terms are included). To determine if the covariation between a daily state and negative events is stronger than between positive events and a daily state, γ_{20} and γ_{10} can be compared using the chi-square technique mentioned above. See Nezlek and Plesko (2001) for examples of such tests.

The moderating effects of individual differences on reactivity to events can also be examined at Level 2. For example, the moderating effects of neuroticism on within-person relationships could be examined with the following model:

$$\text{Intercept: } _{0j} = \gamma_{00} + \gamma_{01}(\text{NEURO}) + u_{0j};$$

$$\text{Positive events: } \beta_{1j} = \gamma_{10} + \gamma_{11}(\text{NEURO}) + u_{1j};$$

$$\text{Negative events: } \beta_{2j} = \gamma_{20} + \gamma_{21}(\text{NEURO}) + u_{2j}.$$

The significance of the γ_{11} and γ_{21} coefficients indicates if neuroticism is related to reactivity to daily events, and differences in the strength of the moderating relationship of neuroticism on reactivity to positive and negative events can be tested by comparing the equality of these coefficients.

Even though the relationship between neuroticism and the intercept may not be of interest in a particular study, neuroticism should be included in the intercept equation because it is likely that average daily state (the intercept) and neuroticism are related. Such inclusion is advisable because MRCM programs use covariance matrices to estimate parameters and test coefficients. A model that did not include neuroticism in the intercept equation but did include it in the slope equation assumes (probably incorrectly) that neuroticism and the intercept are not related. Including neuroticism in the intercept equation increases the accuracy of the parameterization of the covariance matrix and increases the accuracy of estimated parameters and tests of significance. In general, analysts should probably include the same Level 2 variables in each equation and delete those that are not significant, keeping in mind that deleting a variable from one Level 2 equation in a model can change the results of tests of coefficients in other equations of a model.

Variables can be added to both Level 1 and Level 2 models for a variety of purposes: to examine mediational hypotheses, to control for the covariation between measures, and so forth. When deciding on what variables to include in their models, analysts must be mindful of the fact that adding a single variable at Level 1 usually requires the estimation of more than a single additional parameter. Including a variable means that covariances between this and other random errors will also have to be estimated. For example, an analysis in which negative affect (NA) is modeled as a function of positive and negative events estimates three error covariances: the covariance between the intercept and positive events, between the intercept and negative events, and between positive and negative events. If positive affect (PA) is added to the right-hand side of the equation, six covariances are modeled. This includes the same three covariances as in the previous model plus the covariances between PA and the intercept, positive events, and negative event. Of course, if an additional variable is modeled as fixed, then no additional error covariances are estimated.

EVENT-CONTINGENT DATA:
DAY-TO-DAY SOCIAL INTERACTION

Studies of naturally occurring social interaction typify event-contingent methods, and because many studies of naturally occurring social interaction have used variants of the Rochester Interaction Record (RIR) (Wheeler & Nezlek, 1977), the present discussion of the analysis of event-contingent data focuses on the analysis of RIR data. Participants in such studies use standardized forms to describe the social interactions they have each day, and these descriptions typically include the date, time of onset, and duration of the interaction, descriptions of cointeractants such as gender, initials, and relationship with the participant, and various ratings of the interaction.

Social interaction diary data constitute nested data in that observations at one level (interactions) are nested within observations at another level (persons). Hypotheses of interest typically concern interaction quantity (number of interactions per day or percentage of interactions of different kinds) and quality (ratings of interactions such as enjoyment, intimacy, and influence). Hypotheses can concern individual differences in quantity and quality, such as sex differences and relationships between personality variables and interaction; differences among or between different types of interactions, such as interactions with friends and lovers (e.g., Nezlek, 1995) or those involving lies and those that do not (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996); and cross-level effects, such as individual differences in within-subject effects (e.g., Nezlek, Imbrie, & Shean, 1994; Wheeler & Nezlek, 1977). Measures of interaction quantity, such as time spent per day in interaction, can be analyzed as interval-contingent data with days nested within individuals (Nezlek, Hampton, & Shean, 2000), or interaction quantity can be analyzed at the interaction level (percentage of interactions that are a certain type) using techniques appropriate for categorical-dependent measures (Nezlek, 1999).

Although the basic models for these analyses are structurally similar to the models presented in the discussion of daily events, differences between the two types of studies in the nature of the hypotheses of interest require some changes in procedure. Studies of daily events focus primarily on the covariation between events and other measures, whereas with a few exceptions (e.g., Coté & Moskowitz, 1998), studies of social interaction focus primarily on means, with an emphasis on analyzing means representing different types of interactions.

COMPARING TYPES OF INTERACTIONS

Comparing different types of interactions when interactions are classified using a mutually exclusive (nonoverlapping) system is fairly straightforward. For example, many RIR studies have used collegiate samples, and some of these studies have focused on differences among interactions involving only same sex others, only opposite sex others, and interactions with both sexes present (mixed sex). These three categories are nonoverlapping in that an interaction is same, opposite, or mixed sex; it cannot be any two of these simultaneously.

When nonoverlapping classification systems are used, ratings of interactions can be analyzed with a Level 1 model such as

$$y_{ij} = \beta_{1j}(\text{Same}) + \beta_{2j}(\text{Opposite}) + \beta_{3j}(\text{Mixed}) + r_{ij}\,.$$

Each interaction is modeled as a function of three dummy-coded variables (0 for absent, 1 for present) representing each type of interaction. The above model is a zero (or no) intercept model. Using a no intercept model with three dummy-coded variables provides the same tests as a model with a noncentered intercept and two variables; however, the labeling of the output is clearer.[3] To ease the interpretation of coefficients, these variables are not centered; otherwise, estimating mean values for a sample needs to include the distribution of interaction types.

Coefficients (although technically slopes, they represent means) are then passed up to other levels of analysis as in previous examples. Individual differences in these means and their relationships to other individual differences can be analyzed at Level 2 by comparing coefficients in the Level 2 equations, as explained previously. See Nezlek (1999) for an application of this strategy. For example, to compare mean intimacy in same and opposite-sex interactions, the γ_{10} and γ_{20} coefficients from the following model could be compared:

$$\text{Same sex: } {}_{0j} = \gamma_{10} + u_{0j}\,;$$

$$\text{Opposite sex: } \beta_{1j} = \gamma_{20} + u_{1j}\,;$$

$$\text{Mixed sex: } \beta_{2j} = \gamma_{30} + u_{2j}\,.$$

Alternatively, interaction-level contrasts can be created using contrast-coded variables. This allows examination at Level 2 of how the difference between types of interactions varies as a function of individual differences. For example, one variable could compare same- and opposite-sex interactions (same = 1, opposite = –1), and another could compare same- and mixed-sex interactions (same = 1, mixed = –1):

$$y_{ij} = \beta_{0j} + \beta_{1j}(\text{Same-Opp}) + \beta_{2j}(\text{Same-Mix}) + r_{ij}\,.$$

Regardless of whether contrast or dummy codes are used, one caveat is in order about such analyses. If the number of different categories is large and the number of coded variables entered at Level 1 is therefore large, the resulting covariance matrix may be too large to be estimated properly, which is an issue discussed later in the section on fixed and random effects.

MISSING DATA AND ANALYSES OF INTERACTIONS WITH RELATIONAL PARTNERS

Analyses based on overlapping categories are not as straightforward as analyses using nonoverlapping categories, and appropriate analyses vary as a function of the distribution of missing observations. Do all participants have a sufficient number of all types of interactions, and if not, how many participants are missing how many? The primary consideration is the meaningfulness of the pattern of missing data. If there are no missing data or if the pattern of missing data is not meaningful, models (using dummy or contrast codes) similar to those used for nonoverlapping categories can be used. In contrast, if the pattern of missing data is meaningful, a different strategy may be appropriate, and this issue is discussed in terms of the analysis of interactions with relational partners, a variety of whom may be present during any single interaction.

In MRCM analyses, all Level 1 observations contribute to the estimation of parameters regardless of the within-group variability of independent variables.[4] For example, assume a study classifies interactions as a function of the presence or absence of various relational partners (romantic partner, friends, etc.) and assume a substantial number of participants do not have any interactions with one or more of these partners (e.g., romantic partners). One strategy would entail including all interactions for all participants regardless of the within-person distribution of the presence of relational partners. Although this strategy includes the maximum amount of data (all interactions), these analyses could provide misleading parameter estimates because participants who did not have any interactions with a romantic partner would contribute (perhaps substantially) to estimates of parameters describing interactions with romantic partners.

One solution to this dilemma is to conduct separate sets of analyses on data sets consisting only of people who have had interactions with a certain relational partner (or set of partners). For example, analyses of interactions with romantic partners would be done using a data set that contained data only from participants who had a romantic partner. Although this may limit one's ability to compare (statistically) coefficients describing interactions with different relational partners, the coefficients that are estimated are based on the interactions of the people about whom one is making an inference (i.e., people who have such relationships). An example of such an approach can be found in Nezlek et al. (2000).

Another solution involves modeling Level 1 coefficients (intercepts or slopes) at Level 2 as a function of the existence of certain types of missing data. For example, Level 1 coefficients could be analyzed at Level 2 as a function of whether a person had a romantic partner. If coefficients vary as a function of whether a person had a romantic partner, separate analyses as described above may be appropriate, or the existence of a romantic partner can be incorporated in Level 2 analyses.

Large amounts of missing data present complex challenges and difficulties, and the solutions described here are not intended to be prescriptive; rather, they are intended to illustrate possible solutions. Data structures and targets of inference vary considerably across studies, and analysts will have to make decisions about handling missing data on a case-by-case basis. Analysts need to be mindful of the fact that unlike OLS analyses, in which cases with missing observations are generally excluded from analyses, most MRCM programs permit missing data at Level 1. This means that participants with substantial amounts of missing data (either systematically missing or not) will contribute to the estimation of parameters, and the appropriateness of such contributions may vary considerably from situation to situation.

STANDARDIZING VARIABLES

Multilevel random coefficient analyses, like most other random coefficient procedures, rely on covariance, not correlation matrices to estimate parameters. Within the more familiar terminology of OLS analyses, MRCM analyses estimate unstandardized rather than standardized coefficients, and this has important implications for how constructs are measured and how data are prepared for analysis. The focus of MRCM on unstandardized coefficients probably reflects two concerns. First, covariance matrices have more information than correlation matrices. Means and standard deviations are represented in covariance matrices, whereas they are not represented in correlation matrices. Second, MRCM procedures were developed by analysts for whom natural metrics were important. For example, HLM was initially used to analyze educational data, a domain in which questions of interest tend to be phrased in more absolute metrics such as changes in reading levels.

In contrast, for social and personality psychologists, original metrics are frequently not important. Differences in the number of points on scales may reflect differences in researchers' preferences more than differences in the constructs being measured. Consistent with this, social and personality psychologists tend to be interested more in standardized than in unstandardized coefficients. Unfortunately for such researchers, standardizing measures in MRCM analyses is not as straightforward as in OLS-based regression, because the effects of standardization can vary considerably across levels of analysis.

At the macro level of analysis (Level 2 for present purposes), standardization has no implications for significance tests of individual coefficients because changing the variance of a macro-level variable also changes the standard error used to test the significance of the fixed effect. In contrast, standardizing macro-level variables has important implications for testing differences between macro-level coefficients. The importance of metrics for such comparisons is illustrated by analysis from the test data set. Recall that when the intercept for state self-esteem was modeled as a function of adjustment and need for control, a comparison of the coefficients for ADJ and NFC indicated that they were significantly different, $\chi^2(1) = 12.72$. The test data set includes the trait variable ADJ and this same variable multiplied by 10, ADJ10. When the intercept for state self-esteem is modeled as a function of ADJ10 and NFC, a comparison of these coefficients indicates that they are not significantly different, $\chi^2(1) = 1.41$, $p = .23$. Standardizing macro-level variables will have a greater impact on the results of analyses when differences in the variance of raw scores are bigger. Assuming that differences in variances of raw scores are not meaningful, macro-level variables should be standardized if analysts want to compare coefficients involving different macro-level variables.

Decisions about standardizing variables at micro levels (Level 1 for present purposes) are more complicated. First, standardizing Level 1 variables changes Level 1 coefficients and can change the results of significance tests involving these variables. Second, standardizing variables eliminates natural metrics, which may be important. For example, studies of daily events use natural metrics, such as the number of events occurring each day. Such use presumes that there are important differences between days on which many negative events occur and those on which few negative events occur, irrespective of the general distribution of events in a person's life. Ten events are 10 events, and within-person standardization of event scores would eliminate this absolute metric.[5] Nevertheless, researchers may conclude that standardizing Level 1 variables is appropriate; however, such decisions need to be made with the understanding that such standardization will change coefficients and significance tests at all levels of analysis. Finally, researchers who prefer standardized variables at micro levels will probably find it easier to design instrumentation in anticipation of unstandardized analyses than to standardize measures after the fact. One way to maximize the likelihood that a set of Level 1 variables will have similar variances is to use similar if not identical response scales for variables in the set.

DECISION RULES FOR MODELING FIXED AND RANDOM EFFECTS

In most programs that can conduct MRCM analyses, coefficients can be modeled as either random or fixed, and because fixing an effect can have a dramatic influence on the results of analyses (including significance tests), analysts need to make informed choices about fixing effects. In this section, three bases (theoretical, statistical, and practical) for making decisions about whether to model coefficients as fixed or random are discussed.

Although given the focus of most studies, it is probably more appropriate to model interaction- and day-level coefficients as random, coefficients may be fixed for theoretical reasons. For example, assume a researcher believes that the 30 days following the birth of a woman's first child are a critical time. A coefficient describing mothers' reactions across those 30 days could be fixed on the grounds that these 30 days are unique; that is, they are not sampled from a population of days. Nonetheless, researchers need to make strong cases for fixing effects a priori.

Although it may be difficult to make a priori cases for fixing effects, often effects can and should be fixed for statistical reasons. Most MRCM programs provide tests of the significance of the random error for an effect, and nonsignificant random errors should be deleted from a model. Even if a specific random error is significantly different from 0, it may still be statistically justifiable and desirable to delete this parameter from a model. Most MRCM programs estimate covariances among all random effects, and each of these covariances is a parameter (1 *df*). Therefore, when an effect is specified as random, the random variance of the effect and its covariance with other random effects are estimated, and when a random effect is deleted, these covariances are not estimated.

The impact of deleting a particular random effect from a model can be evaluated by examining changes in the overall fit of the model, which is a measure referred to as the deviance statistic. When the deletion of a random parameter produces a nonsignificant change in this statistic, the parameter can and generally should be deleted on the grounds of parsimony. Frequently, effects that are not reliable (below .20) or whose individual error variance is not highly significant can be deleted on this basis. It is important to note that except for rounding error, deviance statistics do not vary as a function of number of iterations. Particularly for complicated models, analysts can sometimes determine which effects can be fixed without having to have a model converge. One advantage of Mln (Rabash, Yang, Woodhouse, &

Goldstein, 1995) compared with HLM is the ability to fix individual elements of covariance matrices. In MLn, the covariation between two errors can be fixed (to be 0 or any value) while individual error terms are left to vary freely. In HLM, the only option is to fix a parameter or not. When a parameter is fixed, the error variance for that term is not estimated nor are the covariances between this error and other error variances.

Finally, as a practical matter, certain coefficients may need to be fixed to get models to converge. Although firm guidelines do not exist, models taking more than 100 to 150 iterations to converge may be misspecified. Frequently, this occurs because a model has exceeded what has been called the carrying capacity of the data. There are too few data points to estimate the parameters of a model. If no clear statistical justification is available, analysts will need to rely on their judgment regarding which random parameters and their covariances are more important. For example, in terms of the constructs being measured, if a parameter covaries only with parameters with which it would not be expected to covary, eliminating such a random parameter would be preferable to eliminating a parameter that covaried more sensibly.

Analysts who have trouble getting models to converge should also consider relaxing the criterion used to determine when a model has converged. For example, in HLM, the default criterion is a .000001 change in the likelihood function, and many models will have likelihood functions that start and finish well into the 1,000s. The effects on parameter estimates and significance tests of relaxing the convergence criterion (e.g., by a factor of 10 or 100) can be determined by examining changes in estimates and tests across models based on different numbers of iterations. If critical parameter estimates and significance tests do not vary across 50, 100, 150, and so on iterations, and the convergence criterion is not fluctuating (which would suggest problems with the stability of covariance matrices), then the convergence criterion can probably be relaxed (in effect, reducing the number of iterations) without affecting the results meaningfully. It is probably preferable to have slightly less accurate estimates of a properly parameterized model (i.e., all effects that should be modeled as random are) than to have rigorously accurate estimates of an improperly parameterized model, particularly given the possibility that fixing effects changes the results of significance tests.

ADDITIONAL CONSIDERATIONS

Given the limitations of a single journal article, it was not possible to discuss in detail various issues that can arise when conducting multilevel analyses. For example, there are specific techniques that need to be used when analyzing categorical dependent measures (the multilevel equivalents of logistic regression), and such techniques are discussed in Raudenbush et al. (2000) and Snijders and Boskers (1999). Determining the power of multilevel designs is not fully understood, although excellent discussions can be found in Snijders and Boskers (1999), Kreft and de Leeuw (1998), and Raudenbush and Liu (1999). Moreover, the program PINT, based on a paper by Snijders and Bosker (1993), can be used to calculate sample sizes for certain two-level designs. Frequently, researchers are concerned about temporal trends in intensive repeated measures data structures such as those discussed here. Although MRCM offers no panacea per se for such problems, most of the types of analyses that have been offered as solutions to such problems (West & Hepworth, 1991) can be accommodated within an MRCM framework. Moreover, multilevel data structures are not immune to problems occurring when assumptions of normality and homogeneity are violated. Although the impact of such violations is not fully understood, preliminary analyses suggest that such violations lead to less efficient (i.e., less powerful) estimates of coefficients (e.g., Bryk & Raudenbush, 1992, chap. 9).

Finally, there is the issue of additional levels of nesting. For the types of analyses described here, an additional (third) level of nesting can be used to conduct multilevel factor and latent variable analyses, and such analyses are described in Bryk and Raudenbush (1992, pp. 191-195) and in Snijders and Boskers (1999). Such analyses have always been available in MRCM programs as a variation of a general procedure, but they are included as a specific procedure in HLM Version 5. Such analyses allow for the examination of multiple dependent variables simultaneously, including comparisons of the strength of slopes across different variables. In addition, for such multivariate analyses, HLM5 allows for examination of specific error structures (Raudenbush et al., 2000, pp. 165-213).

For example, for a daily-events study, items on a scale such as the PANAS could be nested within days and days nested within people. If daily PA and NA are modeled as latent variables, the relationship between PA and some day-level variable such as positive events could be compared with the relationship between NA and the same variable. An additional level of nesting is often needed when hypotheses concern differences in slopes. For example, examining if the relationship (slope) between enjoyment and intimacy in an interaction varies across different types of interactions requires calculating a slope for each type of interaction and then comparing them. This requires nesting interactions within type of interaction and then nesting type of interaction within people.

RECOMMENDATIONS

Social and personality psychologists examine a wide variety of hypotheses and collect a variety of data to examine these hypotheses, making it difficult to provide blanket recommendations about how multilevel data should be analyzed aside from a general recommendation to use some form of random coefficient modeling. Although it is not possible at this time to recommend one technique strongly over the others, two specific techniques, HLM (Raudenbush et al., 2000) and MLn (Rabash et al., 1995), and various procedures in SAS (particularly PROC MIXED) (Littell et al., 1996; Singer, 1998) appear to be the subject of the most attention. It is possible that a consensus may come to exist about the circumstance under which different features of these techniques (and others) represent best practice. See Volume 20 (in 1995) of the *Journal of Educational and Behavioral Statistics* for a discussion of different ways of conducting multilevel analyses.

Although there may be no consensus regarding the techniques constituting best practice, we can increase our understanding of multilevel analyses and the phenomena they describe if reports of multilevel analyses share features. At a minimum, the following aspects of analyses should be reported. Including these details should take no more than a few sentences in any article.

1. Given the diversity of techniques available, authors should avoid potentially confusing references to specific techniques such as HLM (the technique described in Bryk & Raudenbush, 1992) when conducting other types of multilevel analyses. In particular, authors should avoid describing analyses that do not model random coefficients in ways that may lead readers to believe that random coefficients were modeled. For example, readers may be misled by authors who mention that they analyzed their data using a hierarchical linear modeling approach (or words to that effect), cite Bryk and Raudenbush (1992), and then present analyses such as WLS that rely on a qualitatively different statistical model than the analyses presented in Bryk and Raudenbush.
2. It may be useful for authors to describe MRCM analyses with a standard nomenclature. For example, in the multilevel literature, different levels are generally referred to as Level 1, Level 2, and so forth, with an occasional reference to micro and macro levels. The use of terms such as *upper* and *lower* levels or *first* and *second* levels may confuse more than it clarifies. Similarly, authors should try to use standard notation for terms in equations. For example, for two-level models, it is standard practice for coefficients on the right-hand side of Level 1 equations to be represented with the Greek letter β (with subscripts) and for coefficients on the right-hand side of Level 2 equations to represented with the Greek letter γ (with subscripts). The use of other letters such as b and g may confuse more than it clarifies.
3. Given the impact that centering options can have on parameter estimates, researchers should describe what centering options were used. This is particularly important for micro levels of models (e.g., Level 1 of a two-level model) because centering at micro levels influences parameter estimates at macro levels.
4. Given the impact that modeling the same coefficient with or without a random error component can have on parameter estimates, including significance tests of coefficients, researchers should describe which coefficients were modeled as random and which were modeled as fixed. Moreover, researchers should present a justification (theoretical, statistical, or practical) when modeling coefficients without random error components. Similarly, given the general superiority of MRCM over multilevel analyses that do not model random coefficients, authors who use such techniques should provide some justification for this use.

Multilevel data structures are pervasive. As Kreft and de Leeuw (1998) noted, “Once you know that hierarchies exist, you see them everywhere” (p. 1). Nevertheless, although recognizing hierarchies is important, researchers need to know how to analyze hierarchically organized data. This article has presented some of the basic principles of analyses that can expand considerably the hypotheses that social and personality psychologists can examine and can improve substantially the accuracy with which many hypotheses are tested. Although some of the particulars of these analyses are different from the particulars of the ANOVA and regression analyses that constitute the standard canon, their logic is the same. How do observations, be they people, days, or interactions, differ, and how can such differences be explained? MRCM analyses provide powerful tools to answer such questions, and the new generation of software makes such tools accessible to a broad audience. All that remains is for researchers to take advantage of these tools.

APPENDIX A
Observation-Level (Level 1) Data

		Observation									
Subject	*Measure*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	SE	1	1	2	2	3	3	4	4	5	5
	BL	2	3	4	5	6	4	5	7	8	9
B	SE	2	2	3	3	4	4	5	5	4	5
	BL	2	3	4	5	4	5	5	6	7	8
C	SE	3	3	4	4	5	5	5	6		
	BL	3	4	5	6	6	7	8	9		
D	SE	4	4	5	5	6	6	7	7	8	9
	BL	5	4	5	3	3	7	5	6	4	5
E	SE	5	5	6	6	6	7	8			
	BL	3	7	7	8	9	8	8			
F	SE	6	6	6	7	7	8	8	9		
	BL	6	7	8	7	3	5	4	6		
G	SE	7	7	7	8	8	8	9	9	9	
	BL	7	8	6	3	5	8	4	6	9	

(continued)

APPENDIX A Continued

		Observation									
Subject	*Measure*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
H	SE	1	1	2	2	3	3	4	4	5	5
	BL	2	3	4	5	6	4	5	7	8	9
I	SE	2	2	3	3	4	4	5	5	4	5
	BL	2	3	4	5	4	5	5	6	7	8
J	SE	3	3	4	4	5	5	5	6		
	BL	3	4	5	6	6	7	8	9		
K	SE	4	4	5	5	6	6	7	7	8	9
	BL	5	4	5	6	6	7	7	8	8	9
L	SE	7	7	7	8	8	8	9	9	9	
	BL	7	8	6	3	5	8	4	6	9	
M	SE	1	1	2	2	3	3	4	4	5	5
	BL	2	3	4	5	6	4	5	7	8	9
N	SE	2	2	3	3	4	4	5	5	4	5
	BL	2	3	4	5	4	5	5	6	7	8
O	SE	3	3	4	4	5	5	5	6		
	BL	3	4	5	6	6	7	8	9		

APPENDIX B
Person-Level (Level 2) Data

Subject	*ADJ*	*NFC*	*ADJ*10*
A	10	90	100
B	11	87	110
C	12	88	120
D	13	82	130
E	14	86	140
F	15	65	150
G	12	87	120
H	10	87	100
I	11	80	110
J	12	86	120
K	13	83	130
L	12	82	120
M	10	83	100
N	11	85	110
O	12	85	120

NOTE: ADJ = adjustment; NFC = need for control.

NOTES

1. The variability of Level 1 coefficient (slope) can be analyzed whether the coefficient is modeled as random or fixed. When the variability of a fixed coefficient is modeled, this is frequently referred to as an analysis of a non–randomly varying slope or intercept, whereas when the variability of a random coefficient is modeled, the analyses are referred to as randomly varying slopes or intercepts.

2. It is possible to analyze the variance in a slope even when the fixed effect (the mean slope) is not significantly different from 0. For example, if half the Level 2 units had positive slopes and half negative slopes, the mean slope might be 0, but there might be meaningful differences between the Level 2 units that had positive slopes and those that had negative slopes.

3. The equivalence of these two analyses can be seen from the following example. Assume a three-category system in which the data are analyzed with an uncentered intercept and two dummy-coded variables. To interpret the output, predicted scores are estimated for each category. If the dummy codes for Categories 2 and 3 are set to zero, then the intercept represents the coefficient for Category 1, in this case the mean for Category 1. In a no-intercept model, if the dummy codes for Categories 2 and 3 are set to zero, the coefficient for Category 1 also represents the mean for Category 1, the predicted score when Categories 2 and 3 are zero.

4. The output of HLM is potentially misleading in this regard. Chi-square tests of the random error associated with an effect do not include units of analysis for which there is insufficient variability on one or more measures. Nevertheless, all units contribute to the estimation of the fixed effects of coefficients.

5. Micro-level variables can also be standardized in reference to the entire sample of micro-level observations, although such standardization confounds within- and between-level variation. Analysts who standardize in reference to the entire population at the micro-level may want to consider other procedures.

REFERENCES

Affleck, G., Tennen, H., Urrows, S., & Higgins, P. (1994). Person and contextual features of daily stress reactivity: Individual differences of undesirable daily events with mood disturbance and chronic pain intensity. *Journal of Personality and Social Psychology, 66*, 329-340.

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.

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

Barnett, R. C., Marshall, N. L., Raudenbush, S. W., & Brennan, R. T. (1993). Gender and the relationship between job experiences and psychological distress: A study of dual-earner couples. *Journal of Personality and Social Psychology, 64*, 794-806.

Bolger, N., & Zuckerman, A. (1995). A framework for studying personality in the stress process. *Journal of Personality and Social Psychology, 69*, 890-902.

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

Clark, L. A., & Watson, D. (1988). Mood and the mundane: Relations between daily life events and self-reported mood. *Journal of Personality and Social Psychology, 54*, 296-308.

Coté, S., & Moskowitz, D. S. (1998). On the dynamic covariation between interpersonal behavior and affect: Prediction from neuroticism, extraversion, and agreeableness. *Journal of Personality and Social Psychology, 75*, 1032-1046.

Cutrona, C. E. (1986). Behavioral manifestations of social support: A microanalytic investigation. *Journal of Personality and Social Psychology, 51*, 201-208.

de Leeuw, J., & Kreft, I. G. G. (1995). Questioning multilevel models. *Journal of Educational and Behavioral Statistics, 20*, 171-189.

DePaulo, B., Kashy, D. A., Kirkendol, S. A., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70*, 979-995.

Gable, S. L., & Reis, H. T. (1999). Now and then, them and us, this and that: Studying relationships across time, partner, context, and person. *Personal Relationships, 6*, 415-432.

Hodgins, H. S., Koestner, R., & Duncan, N. (1996). On the compatibility of autonomy and relatedness. *Personality and Social Psychology Bulletin, 22*, 227-237.

Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., Vol. 1, pp. 233-265). New York: McGraw-Hill.

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

Kreft, I.G.G., de Leeuw, J., & Aiken, L. (1995). The effects of different forms of centering on hierarchical linear models. *Multivariate Behavioral Research, 30*, 1-22.

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. (1995). Social construction, gender/sex similarity, and social interaction in close personal relationships. *Journal of Social and Personal Relationships*, *12*, 503-520.

Nezlek, J. B. (1999). Body self-evaluation and day-to-day social interaction. *Journal of Personality*, *67*, 793-817.

Nezlek, J. B., Hampton, C., & Shean, G. D. (2000). Clinical depression and everyday social interaction in a community sample. *Journal of Abnormal Psychology*, *107*, 11-19.

Nezlek, J. B., Imbrie, M., & Shean, G. D. (1994). Depression and everyday social interaction. *Journal of Personality and Social Psychology*, *67*, 1101-1111.

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., & Zyzniewski, L. E. (1998). Using hierarchical linear modeling to analyze grouped data. *Group Dynamics: Theory, Research, and Practice*, *2*, 313-320.

Rabash, J., Yang, M., Woodhouse, G., & Goldstein, H. (1995). *MLn: Command reference guide*. London: Institute of Education.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2000). *HLM5*. Chicago, IL: Scientific Software.

Raudenbush, S. W., & Liu, X. (1999). *Statistical power and optimal design for multisite randomized trials.* Ann Arbor: University of Michigan, Survey Research Center.

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. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237-259.

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

Suls, J., Green, P., & Hillis, S. (1998). Emotional reactivity to everyday problems, affective inertia, and neuroticism. *Personality and Social Psychology Bulletin*, *24*, 127-136.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality*, *59*, 611-662.

Wheeler, L., & Nezlek, J. (1977). Sex differences in social participation. *Journal of Personality and Social Psychology*, *35*, 742-754.

Received March 10, 1999

Revision accepted June 16, 2000