Multilevel Modeling: Current and Future Applications in Personality Research

Stephen G. West,[1] Ehri Ryu,[2] Oi-Man Kwok,[3] and Heining Cham[1]

[1]Arizona State University

[2]Boston College

[3]Texas A&M University

ABSTRACT Traditional statistical analyses can be compromised when data are collected from groups or multiple observations are collected from individuals. We present an introduction to multilevel models designed to address dependency in data. We review current use of multilevel modeling in 3 personality journals showing use concentrated in the 2 areas of experience sampling and longitudinal growth. Using an empirical example, we illustrate specification and interpretation of the results of series of models as predictor variables are introduced at Levels 1 and 2. Attention is given to possible trends and cycles in longitudinal data and to different forms of centering. We consider issues that may arise in estimation, model comparison, model evaluation, and data evaluation (outliers), highlighting similarities to and differences from standard regression approaches. Finally, we consider newer developments, including 3-level models, cross-classified models, nonstandard (limited) dependent variables, multilevel structural equation modeling, and nonlinear growth. Multilevel approaches both address traditional problems of dependency in data and provide personality researchers with the opportunity to ask new questions of their data.

Generations of personality researchers were taught to avoid dependency in data. To use the common tools of statistical analysis in

The online version of this article contains additional supporting information, including (a) a more detailed description of the method used to simulate the data presented in this article and the corresponding SAS computer script and (b) SAS and SPSS computer script for all analyses reported in this article.

Correspondence concerning this article should be addressed to Stephen G. West, Psychology Department, Arizona State University, Tempe, AZ 85287-1104. Email: sgwest@asu.edu.

Journal of Personality 79:1, February 2011

DOI: 10.1111/j.1467-6494.2010.00681.x

personality psychology—analysis of variance, multiple regression, factor analysis—authors emphasized that observations must be independent or that serious inflation of the Type 1 error rate would occur, making the results untrustworthy. Simple statistical solutions existed when data could be highly structured, as in the classic laboratory Person × Situation design in which all persons could be measured in all situations (Krahé, 1992; Snyder & Ickes, 1985)—with situations typically presented in the form of hypothetical vignettes. But the lack of good statistical solutions discouraged researchers from studying interactions in more naturalistic social environments. Studies of the effects of natural groups such as families or friendship groups on the personality of individual members (and vice versa) also lagged. Researchers attempting to analyze data on these issues using the commonly available statistical tools found themselves chastised by reviewers for their failure to address fundamental issues of dependency in their data. Similar data analytic complexities arose for longitudinal researchers given the inherent dependency of data collected from the same individuals at multiple time points, particularly when the data could not be measured at a fixed set of common measurement points, the theoretical questions could not be answered by repeated measures analysis of variance, or both.

Versatile statistical tools for addressing dependency in data have been developed over the past 25 years. These tools have the remarkable property that they not only can address problems of dependency (clustering) in data, but they also can allow us to exploit the dependency, allowing researchers to ask important new questions of their data. Observations may be clustered within experimental treatments (e.g., patients within group treatment conditions), or clustered within natural groups (e.g., students within classrooms), or repeated measures may be clustered within the individual from whom they were collected. One of the most important of these approaches is multilevel modeling (MLM). MLM provides proper parameter estimates and standard errors for clustered data. MLM also capitalizes on the hierarchical structure of the data, allowing researchers to study relations among variables at different levels and even across levels.

The focus of this article will be on multilevel extensions of univariate analysis methods such as multiple regression (e.g., hierarchical linear modeling, mixed modeling). These approaches originated in the work of Charles Henderson (1953) in agricultural statistics, were

developed in basic and applied areas of statistics, and have seen application in many research areas (Bryk & Raudenbush, 1992; Gelman & Hill, 2007; Goldstein, 1995; Hox, 2010; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Multilevel extensions of multivariate techniques such as multilevel structural equation modeling (Goldstein & McDonald, 1988; Lee, 1990; Longford & Muthén, 1992; Muthén & Satorra, 1995) and longitudinal models of growth and change (Biesanz, West, & Kwok, 2003; Bollen & Curran, 2006; McArdle, 2009; Singer & Willett, 2003) have also been developed. Given space constraints, more complex multivariate and longitudinal extensions can only be briefly considered here.

In MLM, data are structured in two or more hierarchical levels. An example of a three-level structure is repeated observations (Level 1) within children (Level 2) who are in turn measured within school classrooms (Level 3). In most applications, data are available at only two of the levels (e.g., repeated observations within persons; persons within school classrooms). Additional hierarchical levels can potentially be identified, but only those at which data are collected or at which nontrivial dependency is expected to occur need be considered in the statistical analysis. Predictor variables can be included (or not) at each level of the hierarchy and the interactions between predictors at different levels can be considered. When time-related variables are considered in the model, trajectories of the growth or decline of personality traits can be estimated and individual differences in those trajectories studied (Singer & Willett, 2003).

Beyond the potential usefulness of these statistical tools, they are becoming increasingly available. The basics of the statistical theory are increasingly being taught in graduate training programs in psychology (Aiken, West, & Millsap, 2008), good readable introductory texts (e.g., Hox, 2010; Snijders & Bosker, 1999) and advanced resources (e.g., de Leeuw & Meijer, 2008; Gelman & Hill, 2007; Raudenbush & Bryk, 2002) are now available, and user-friendly software routines in standard (e.g., SAS, SPSS,[1] STATA) and specialized statistical packages (e.g., GLLAMM, HLM, MLwiN, Mplus) are available. Reflecting its development in diverse

1. SPSS has been acquired by IBM. It is unknown as of this writing how the performance of the SPSS programs will be affected, if at all, by this change. We will refer to the IBM SPSS software as SPSS in this article.

applied areas of statistics, MLM is called by many names, with multilevel modeling, hierarchical linear modeling, mixed models, and random coefficient models being a few of the more prominent ones.

We initially review the current use of MLM in personality research. As we will see below, personality researchers are using MLM in their research, but its current use appears to be primarily limited to the two research domains of experience sampling and studies of longitudinal growth. Following in the tradition of previous introductions (Cohen, Cohen, West, & Aiken, 2003, chap. 14; Nezlek, 2001, 2007), in the first half of the article we present an introduction to the basics of standard MLM, focusing on its current use in personality research. This section will be particularly useful for personality researchers who wish to begin using MLM in their work. In contrast, the second half of the article identifies features of MLM commonly neglected by personality researchers: the realms of estimation, model evaluation, and data evaluation (outliers). This section will be useful even for experienced MLM researchers. Finally, we point to some extensions of multilevel analysis that may have potentially important future applications in personality research, permitting researchers to address new questions.

CHARACTERIZING THE CURRENT STATUS OF MLM IN PERSONALITY RESEARCH

To provide a snapshot of the current use of MLM in mainstream personality research, we conducted a search of the literature for the period January 2007 to June 2008 in three major personality journals[2]: *Journal of Personality* (*JP*), *Journal of Personality and Social Psychology: Personality Process and Individual Differences* (*JPSP: PPID*), and *Journal of Research in Personality* (*JRP*). Based on both a computer search using keywords (e.g., *multilevel model*) and a

2. Given our goal of characterizing mainstream personality research, we did not consider articles in journals in abnormal psychology or life span development that included personality variables. Our use of a limited time window also implies that studies that are outliers because they have rare features may not be included because they fall outside the time window. For example, a few personality studies have reported secondary analyses of existing data sets with very large sample sizes (e.g., Lucas, Clark, Georgellis, & Diener, 2003, German Socio-Economic Panel study, $n > 24{,}000$, 15 annual reports).

manual search, we identified a total of 25 regular articles[3] that used multilevel or growth curve models (*JP*: $n = 6$, 10% of total articles; *JPSP: PPID*: $n = 9$, 15% of total articles; *JRP*: $n = 10$, 7% of total articles). We selected the first study in each article that employed multilevel or growth curve analyses for more detailed review. Although a variety of extensions of standard multilevel models are now available to address more than two hierarchical levels in the data and a variety of types of dependent variables (e.g., counts), all articles reviewed considered the standard two-level model. With one exception (binary dependent variable; McLean & Fournier, 2008), the dependent variable was always treated as continuous. As described below, two prototypical forms of data collection were reported; MLM analyses appear to have become the standard expected analyses in these areas, perhaps having become part of the current "craft" research knowledge in those domains.

First, 9 of the 25 studies were classified as experience sampling studies (see Conner, Feldman-Barrett, Tugade, & Tennen, 2007; Reis & Gable, 2000) in which multiple observations were collected on each participant. The typical focus of these studies is on the effect of natural variation in events or situations and their effect on more "statelike" outcome variables (e.g., emotions, happiness). With one exception (Tamir, John, Srivastava, & Gross, 2007), the time ordering of the observations was ignored. An illustration of this type of study is provided by Oishi, Diener, Choi, Kim-Prieto, and Choi (2007), who conducted a 21-day Web-based study of life satisfaction. At Level 1, each participant reported on a daily basis on positive and negative events (predictor variables) and life satisfaction (outcome variable). At Level 2, participants were sampled at universities in three different countries to represent distinct cultural groups (European Americans, Asian Americans, Koreans, Japanese). Across the 13 experience sampling studies we reviewed, the median number of observations per participant was 20 (range = 4–132), and the median number of participants at Level 2 was 113 (range = 26–332).

Second, 9 of the 25 studies were longitudinal growth models in which repeated observations on the same measure were collected on individuals and the goal was to model growth trajectories. To illustrate, Cramer and Jones (2007) analyzed data from measures of

3. One special issue of *Journal of Personality* was not considered. All full-length articles and brief reports were reviewed.

self-control and self-acceptance collected from adults at four measurement waves spanning up to 33 years (Level 1). At the initial measurement, each participant was assessed on the defense measures of denial and identification (Level 2), assumed to be chronic individual differences. For the nine studies of longitudinal growth we reviewed, the median number of Level 1 observations per person was 4 (range = 3–15), whereas the median number of Level 2 observations (persons) was 339 (range = 130–1,692). Of the nine studies, six were analyzed within the MLM framework and three within the structural equation modeling (SEM) framework. For standard applications, MLM and SEM provide identical results (Curran, 2003; Mehta & West, 2000), although one of the approaches may be more appropriate for specific advanced applications (e.g., MLM for three or more level models, SEM for multilevel confirmatory factor analysis models).

Among the remaining studies we reviewed, three were classic multilevel studies that collected data from participants in larger groups (classrooms, day care centers). In two of these studies, no Level 2 variables were collected and MLM was simply used to correct for dependency in the data (see Oishi, Lun, & Sherman, 2007, for the exception). Three other studies collected multiple reports (e.g., autobiographical memories; McLean & Fournier, 2008) from participants. Statistically paralleling the experience sampling studies, the effects of report-based Level 1 predictors (e.g., recall effort) and Level 2 individual difference predictors (e.g., neuroticism) on the outcomes were examined. Finally, one study (Biesanz, West, & Millevoi, 2007) collected data from multiple peers (Level 1) who differed in the length of acquaintance with participants (Level 2). Besides helping complete the portrait of current research designs, these studies begin to indicate the divergent hierarchical data structures to which MLM can be applied, possibly suggesting new applications that can be exploited in the future.

BASICS OF MULTILEVEL MODELING

As we have seen, multilevel data structures can occur in personality research in various ways, such as diary data, longitudinal data, individuals within natural groups, or multiple peer reports. In this section, we introduce the basic ideas of MLM using artificial data.

The data were based on a study by Armeli, Carney, Tennen, Affleck, and O'Neil (2000). These authors were interested in examining the relationship between alcohol use and stress and how other covariates such as alcohol outcome expectancy and gender might moderate these relationships. Based on the original results of Armeli et al. (2000), we simulated daily diary data representing four measures from 100 individuals (50 women and 50 men) for a period of 70 days (10 weeks): alcohol consumption (*ALC*), stress (*STR*), gender (*GEN*), and positive alcohol-outcome expectancy (*AOE*). The dependent variable, alcohol consumption, and a time-varying predictor variable, stress, were measured each day. Two time-invariant predictors (*GEN*, *AOE*) were measured once at the beginning of the study and assumed not to change over time. The means, standard deviations, skewness, and kurtosis of generated variables are reported in Table 1. The models presented here were estimated using the SAS 9.1 Mixed procedure (see Supplement B(1)).

In our presentation below, we will consider a series of models that explore a variety of hypothesized relationships. The series of models is presented in part for the pedagogical reason of illustrating some of the variety of the models that can be estimated. Also of importance, some useful statistics can only be obtained by comparing models—the simpler models necessary for the comparisons are presented here. In addition, as in multiple regression in which the inclusion of additional predictors can affect the results, the comparison of

Table 1
Means, Standard Deviations, Skewnesses, and Kurtoses of *ALC*, *STR*, *GEN*, and *AOE*

Variable	Mean	*SD*	Skewness	Kurtosis
ALC (Alcohol consumption)[a]	2.19	1.12	0.31	−0.18
STR (Stress)[a]	2.01	0.37	0.12	4.15
GEN (Gender)[b]	0.00	0.50	0.00	−2.00
AOE (Positive alcohol-outcome expectancy)[b]	2.32	1.04	0.35	−0.83

[a]Time-varying variable based on $N = 7{,}000$ observations (100 participants × 70 daily reports per participant).
[b]Time-invariant variable based on 100 participants. *SD* = standard deviation. Gender was contrast coded: male = −0.5, female = +0.5.

multiple models can inform the interpretation of the results of the more complex models. We initially consider three models that ignore the effects of the time variable, paralleling this common practice in the experience sampling literature. We then indicate how inclusion of a time-related variable might sometimes add to our understanding of the results. Finally, we show how the use of specific centering procedures can yield results whose interpretation is particularly useful for some research questions. In our presentation, we encourage readers to pay careful attention to the interpretation of effects, which can change in important ways across models. Failure to consider the interpretation of the effects can lead to considerable confusion in the literature because superficially similar effects may be testing distinctly different hypotheses, sometimes leading to dramatically different results (West, Aiken, Wu, & Taylor, 2007). The results for all of the models are presented in Table 2.

Model A: Random Intercept Model

We begin with the simplest model, the random-intercept model (a.k.a. unconditional model) with no predictors. The random-intercept model partitions the variance in the dependent measure into within-individual and between-individuals variance components. Equation (1) represents the within-individual mean and variability in *ALC*. Equation (2) represents the between-individuals mean and variability. For our illustration, the random-intercept model is written as

$$\text{Level 1 (within individual)}: (ALC)_{ti} = \beta_{0i} + e_{ti} \quad (1)$$

$$\text{Level 2 (between individuals)}: \beta_{0i} = \gamma_{00} + u_{0i} \quad (2)$$

Two subscripts are needed to describe each data point. t indicates time (measurement occasions, $t = 1, 2, \ldots, 70$); i indicates individual ($i = 1, 2, \ldots, 100$). In equation (1), β_{0i} represents the mean alcohol consumption for the ith individual across the 70 days of the study. On each day, person i may drink more or less than his own mean level of alcohol consumption. The deviation of each daily measure from the individual's own mean on *ALC* is captured by the Level 1 residual e_{ti}. In standard MLM, two assumptions are commonly made. First, the Level 1 residual e_{ti} is assumed to be normally distributed with mean 0 and variance σ^2. Second, the variance of e_{ti} is assumed to be homogeneous across individuals (i.e., $e_{ti} \sim N(0, \sigma^2)$).

Table 2
Results of Multilevel Modeling Analyses: Five Models

	Model A	Model B	Model C	Model D	Model E
Intercept					
Intercept ($\hat{\gamma}_{00}$)	2.19*	1.69*	1.14*	1.66*	2.19*
Gender ($\hat{\gamma}_{01}$)			−0.20		−0.87*
AOE ($\hat{\gamma}_{02}$)			0.24*		0.49*
Gender × *AOE* ($\hat{\gamma}_{03}$)			−0.01		−0.38*
$\overline{Stress}_{person}$					−0.82*
Slope for stress					
Intercept ($\hat{\gamma}_{10}$)		0.25*	−0.03	0.21*	0.25*
Gender ($\hat{\gamma}_{11}$)			0.10		−0.32*
AOE ($\hat{\gamma}_{12}$)			0.12*		0.12*
Gender × *AOE* ($\hat{\gamma}_{13}$)			−0.18*		−0.18*
$\overline{Stress}_{person}$					−0.41*
Slope for *WKEND*					
Mean ($\hat{\gamma}_{20}$)				0.43*	
Variance estimates					
Level 1 residual ($\hat{\sigma}^2$)	0.42*	0.40*	0.40*	0.34*	0.40*
Level 2 residuals					
Var(u_{0i}), or $\hat{\tau}_{00}$	0.84*	0.25*	0.19*	0.25*	0.36*
Var(u_{1i}), or $\hat{\tau}_{11}$		0.09*	0.05*	0.09*	0.04*
Var(u_{2i}), or $\hat{\tau}_{22}$				0.13*	
Cov(u_{0i}, u_{1i}), or $\hat{\tau}_{10}$		0.05*	0.00	0.07*	0.08*
		($r = .35$)	($r = .00$)	($r = .47$)	($r = .69$)
Cov(u_{0i}, u_{2i}), or $\hat{\tau}_{20}$				−0.12*	
				($r = -.64$)	
Cov(u_{1i}, u_{2i}), or $\hat{\tau}_{21}$				0.01	
				($r = .08$)	

Notes. *AOE* = positive alcohol-outcome expectancy; *Gender* is contrast coded: male = −0.5, female = +0.5; *WKEND* = 0 for Sunday through Thursday, *WKEND* = 1 for Friday and Saturday. In Model E, *Stress* was centered at person mean. $\overline{Stress}_{person}$ = person mean (chronic) level of stress. In Model E, all Level 2 predictors were centered at the grand means.
*$p < .05$.

These assumptions permit us to represent the variation in parameters that occurs at Level 1 with just two parameter values at Level 2, a mean and a variance. In some cases, more complicated representations (e.g., differences in residual variances between subgroups:

male, female) may be necessary to accurately portray the data. Options exist in MLM programs to specify these models.

In Equation (2) γ_{00} is the grand mean level of alcohol consumption in the population of college students from which the researchers selected their sample. To clarify how terms are calculated in the sample, we will use $\wedge$ over the term to represent the sample-based estimate of the population value (e.g., $\hat{\gamma}_{00}$). For the 100 participants, $\hat{\gamma}_{00}$ is the grand mean of their mean drinking level across the 70 days of the study. Individual differences from the grand mean level are captured by the Level 2 residual u_{0i}: how much does person *i*'s mean level of drinking differ from the grand mean level of the full sample? In standard MLM, the Level 2 residuals are assumed to be normally distributed in the population with mean 0 and variance τ_{00} (i.e., $u_{0i} \sim N\,(0, \tau_{00})$).

The estimates are shown in Table 2 in the column for Model A. The mean alcohol consumption $\hat{\gamma}_{00}$ was 2.19 drinks per day. The within-individual (Level 1) variance $\hat{\sigma}^2$ of alcohol consumption was 0.42 ($SD = 0.65$), and the between-individuals (Level 2) variance $\hat{\tau}_{00}$ was 0.84 ($SD = 0.92$). A useful index of the degree of dependency (clustering), the intraclass correlation (ICC), can be computed based on the random intercept model (Model A) using the Level 1 ($\hat{\sigma}^2$) and Level 2 ($\hat{\tau}_{00}$) variance estimates.

$$ICC = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2} \tag{3}$$

One interpretation of the ICC that is useful in many personality research designs is that the ICC is the proportion of the total variance that is accounted for by Level 2 (here, between-individuals) variance. Another interpretation is that it is the extent to which measurements taken on the same cluster (here, person) are more similar than measurements taken on different clusters (persons).[4] The ICC in our illustrative example was $0.84/(0.84+0.42) = 0.67$,

4. Other definitions of the ICC may be appropriate in other contexts. Shrout and Fleiss (1979) presented definitions for the use of the ICC as a measure of interjudge reliability when multiple judges are employed. Kenny, Kashy, and Bolger (1998) described experimental contexts in which similar participants (two friends) are experimentally assigned to be in different dyads, leading to negative ICCs. The ICC only takes into account differences in intercepts between clusters, not differences in slopes (Roberts, 2007). Slope differences are the focus of the MLM models considered below.

which indicates 67% of the total variance in the alcohol consumption is due to between-individuals differences in mean levels. The other 33% of the variance is due to the within-individual variability across the 70 days of the study. When the Level 1 unit is observations within individuals and the Level 2 unit is individuals (as in experience sampling studies), the ICC is expected to be high, here nearly .70. In contrast, in multilevel structures in which the Level 1 unit is individuals within groups and the Level 2 unit is groups, an ICC value of about .20 or .30 is often deemed high.

Model B: Random Coefficient Model With Level 1 Predictor

In Model B, we add a time-varying predictor, stress (STR), to account for ALC at Level 1. The predictor, STR, is included at Level 1 because it was measured on a daily basis and can potentially vary from day to day for any individual. We now need one equation (4) at Level 1 and two equations (5, 6) at Level 2 to represent the model.

$$\text{Level 1 (within individual): } (ALC)_{ti} = \beta_{0i} + \beta_{1i}(STR)_{ti} + e_{ti} \quad (4)$$

$$\text{Level 2 (between individuals): } \beta_{0i} = \gamma_{00} + u_{0i} \quad (5)$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \quad (6)$$

At Level 1, $\hat{\beta}_{0i}$ represents the estimate of the regression intercept for person i. The meaning of this coefficient has changed from Model A. $\hat{\beta}_{0i}$ no longer represents person i's mean level of drinking but rather person i's predicted alcohol consumption *when his or her level of stress is* 0. This change in interpretation reminds us to pay careful attention to the meaning of the value of 0 on each predictor—indeed, 0 may be a value that does not even exist on the scale! Here we will continue to use the original scaling: in this example, $STR = 0$ represents a meaningful value in which no life events have occurred (we consider centering procedures that can help minimize interpretational issues in a later section). $\hat{\beta}_{1i}$ represents the linear relationship (slope) between stress and alcohol consumption for person i. The Level 1 residual e_{ti} is the difference between person i's observed and predicted alcohol consumption on each day of the study. In the Level 1 equation, the subscript i in the notation for the estimates of the regression intercept ($\hat{\beta}_{0i}$) and slope ($\hat{\beta}_{1i}$) reminds us that individuals may differ in their intercepts, slopes, or both.

The Level 2 model (between-individuals model) captures the individual differences in the true levels of β_{0i} and β_{1i} for each individual. γ_{00} and γ_{10} represent the grand means in the population of intercepts and slopes across individuals, respectively. Conceptually, $\hat{\gamma}_{00}$ represents the grand mean of the 100 intercepts ($\hat{\beta}_{0i}$), and $\hat{\gamma}_{10}$ represents the grand mean of the 100 slopes ($\hat{\beta}_{1i}$) that would be estimated for each individual based on his or her 70 days of data. The Level 2 residuals u_{0i} and u_{1i} represent deviations of each individual's intercept and slope from the grand intercept γ_{00} and grand slope γ_{10}, respectively. Given that both slope and intercepts are now being estimated at Level 2, a slightly more complicated assumption is needed. The Level 2 residuals are now assumed to be multivariately normally distributed: $\begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \\ \tau_{10} & \tau_{11} \end{bmatrix}\right)$. This assumption allows us to represent all of the information about individual intercepts and slopes at Level 2 in terms of means, variances, and covariances.

As shown in Table 2 in the column for Model B, the estimated grand mean intercept $\hat{\gamma}_{00} = 1.69$, which indicates the predicted mean alcohol consumption *now when STR is 0*. The estimated grand mean slope $\hat{\gamma}_{10}$ for *STR* was 0.25, which indicates that on average, individuals consumed 0.25 more alcoholic drinks for each 1-unit increase on the stress measure (0–3 scale). The significant intercept variance ($\hat{\tau}_{00} = 0.25$, $SD = 0.50$) indicates that the estimated alcohol consumption at $STR = 0$ differs significantly between individuals.[5] The significant slope variance ($\hat{\tau}_{11} = 0.09$, $SD = 0.30$) indicates that the linear relationship between stress and alcohol use differs from individual to individual. There was also a significant positive covariance between the intercept and slope coefficients for the 100 individuals (covariance $\hat{\tau}_{10} = 0.05$, corresponding to a correlation of +0.35), indicating that those who drink more when $STR = 0$ tend to have larger increases in *ALC* as stress increases (i.e., positive relationship).

5. The Wald test reported for the variance is not optimal in two ways. First, variances can only be 0 or positive, so a one-tailed test is appropriate, implying that the reported two-tailed *p* values should be halved (Snijders & Bosker, 1999, p. 90). Second, the null hypothesis that the variance = 0 represents a boundary condition so that neither the Wald test nor the likelihood ratio test have the appropriate distribution. These issues are considered in a later section on model comparison procedures.

Model C: Random Coefficient Model With Level 1 and Level 2 Predictors

In the next model (Table 2, Model C), we studied the effects of the two time-invariant (Level 2) predictors collected at baseline: gender (*GEN*) and positive alcohol-outcome expectancy (*AOE*). In the Level 2 equations, we included each of these predictors and their interaction to account for individual differences in the intercept for the *STR-ALC* relationship. *GEN* was contrast coded, male = -0.5 and female = $+0.5$, facilitating easier interpretation (see Cohen et al., 2003, chap. 8). Our equations are

$$\text{Level 1 (within individual): } (ALC)_{ti} = \beta_{0i} + \beta_{1i}(STR)_{ti} + e_{ti} \tag{7}$$

$$\text{Level 2 (between individuals): } \beta_{0i} = \gamma_{00} + \gamma_{01}(GEN)_i + \gamma_{02}(AOE)_i + \gamma_{03}(GEN)_i(AOE)_i + u_{0i} \tag{8}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}(GEN)_i + \gamma_{12}(AOE)_i + \gamma_{13}(GEN)_i(AOE)_i + u_{1i} \tag{9}$$

Equation (7) represents the Level 1 model, and the expression and its interpretation are identical to the Level 1 equation (Equation 4) in Model B. Daily alcohol consumption is predicted by daily stress at the within-individual level. Equation (8) is the Level 2 model for the intercept. γ_{00} is the population regression intercept (recall that the regression intercept, $\hat{\beta}_{0i}$, is the predicted alcohol consumption when $STR = 0$) when $AOE = 0$ and $GEN = 0$. Because we used contrast codes to represent *GEN*, GEN is centered (= 0) at the unweighted mean of women and men, permitting us to ignore gender in the interpretation of other lower order effects. γ_{01} is the difference in regression intercepts between men and women when $AOE = 0$. γ_{02} is the linear relationship between the intercept and *AOE*. Finally, γ_{03} is the interaction between *GEN* and *AOE* that may account for differences in the individual intercepts β_{0i}.

Equation (9) is identical in its interpretation to Equation (8), except that the effects being described are based on the slopes rather than the intercepts. γ_{10} is the predicted regression slope (again recall that the regression slope, β_{1i}, is the average relationship between stress and alcohol consumption) when $AOE = 0$ (*GEN* can be ignored since it is centered). γ_{11} is the difference in the slope between men and women when $AOE = 0$. γ_{12} is the relationship between the slope and *AOE*. Finally, γ_{13} is the interaction between *GEN* and

AOE that may account for individual differences in the individual slopes. One difference from Equation (8) is that in Equation (9) γ_{11} and γ_{12} represent the effect of Level 2 predictors on the relationship between stress and alcohol consumption—in other words, the moderating effects of Level 2 predictors. Therefore, γ_{11} and γ_{12} represent two-way *cross-level interactions* between the Level 1 (*STR*) and Level 2 predictors (*GEN*, *AOE*, respectively). Likewise, γ_{13} represents the three-way cross-level interaction of *STR*, *GEN*, and *AOE*. This difference can be seen more clearly if we combine Equations (7), (8), and (9) into a single reduced-form equation (mixed model). The three cross-level interaction terms are now clearly indicated in the second line of Equation (10) by product terms involving the Level 1 predictor *STR* and the Level 2 predictors *GEN* and *AOE*:

$$\begin{aligned}(ALC)_{ti} = {} & \gamma_{00}+\gamma_{10}(STR)_{ti}+\gamma_{01}(GEN)_i+\gamma_{02}(AOE)_i+\gamma_{03}(GEN)_i(AOE)_i \\ & +\gamma_{11}(STR)_{ti}(GEN)_i+\gamma_{12}(STR)_{ti}(AOE)_i+\gamma_{13}(STR)_{ti}(GEN)_i(AOE)_i \\ & + u_{0i} + u_{1i}(STR)_{ti} + e_{ti}\end{aligned} \tag{10}$$

In Table 2 in the column for Model C, the estimated grand intercept $(\hat{\gamma}_{00}) = 1.14$ indicated that the predicted mean alcohol consumption for those who have stress = 0 and positive alcohol outcome-expectancy (*AOE*) = 0 would be 1.14. *AOE* was positively related to the value of the individual intercepts, representing alcohol consumption under zero stress ($\hat{\gamma}_{02} = 0.24$, $p < .05$). *STR* was not significantly related to alcohol consumption when *AOE* was 0 ($\hat{\gamma}_{10} = -0.03$, *ns*). However, the relationship between *STR* and *ALC* was moderated by *AOE*, $\hat{\gamma}_{12} = 0.12$, $p < .05$. As *AOE* increased in value, the relationship of stress to alcohol consumption became increasingly positive. Finally, there was an interaction between *GEN* and *AOE*, $\hat{\gamma}_{13} = -0.18$, $p < .05$, that explained a portion of the individual differences in the slope (the linear relationship between stress and drinking).

We used an extension of Aiken and West's (1991) simple slopes procedure to probe the interaction. The slopes for the relationship between stress and alcohol consumption at three different values of *AOE* (Low: Mean − 1 *SD*, Mean, and High: Mean +1 *SD*) are plotted separately for men and women in Figure 1. As can be seen, the moderating effect of *AOE* on the stress–alcohol consumption relationship appeared to be stronger for men than for women. In fact,

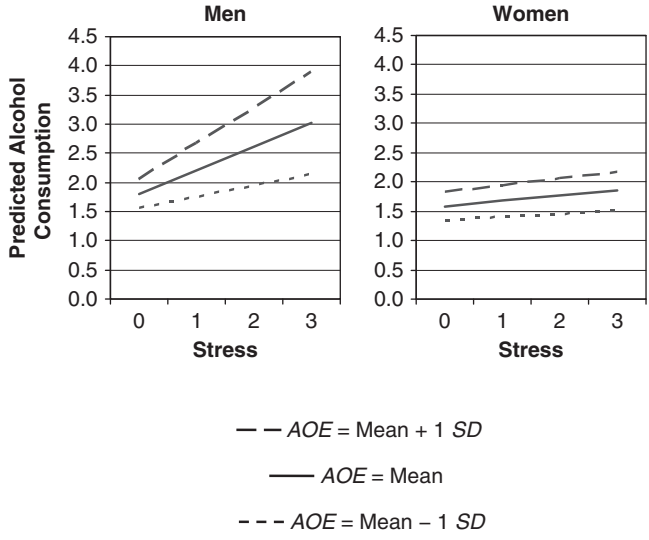

Figure 1
Relationship of stress and alcohol consumption at different levels of *AOE* for women and men. *AOE* = positive alcohol-outcome expectancy.

the moderating effect of *AOE* on the stress-drinking relationship was significant for men but not for women. Bauer and Curran (2005) presented methods for extending the Aiken-West procedure for testing simple slopes in the MLM context. And Preacher, Curran, and Bauer (2006) presented an online Web-based calculator that facilitates the computations. For men, the estimated simple slope of stress was 0.63, $p < .05$ when $AOE = \text{Mean} + 1\ SD$ (high *AOE*), and 0.40, $p < .05$ when $AOE = \text{Mean} - 1\ SD$ (low *AOE*). For women, the estimated simple slope of stress was 0.11, *ns*, and 0.08, *ns*, at high and low *AOE*, respectively.

At a more general level, we can examine the reduction of the Level 2 variances τ_{00} and τ_{11} from Model B to Model C to gain an understanding of the overall effect of the inclusion of the Level 2 predictors (stable individual differences) on the prediction of the individual intercepts and slopes, respectively. The Level 2 predictors explained 24% of the intercept variance ($(0.25 - 0.19)/0.25 = 0.24$) and 44% of the variance in the relationship between stress and alcohol consumption ($(0.09 - 0.05)/0.09 = 0.44$). As noted by Kreft and de Leeuw (1998) and Snijders and Bosker (1999), caution must be exercised in

interpreting variance reduction measures (R^2 analogs) because of the multiple levels and the differences in estimation procedures between multiple regression and MLM. Variance estimates can change in unexpected ways as predictors are added to or deleted from the model. Indeed, the addition of a predictor can in some cases even lead to a decrease in the proportion of variance accounted for, an impossible result in multiple regression. As an alternative approach to computing explained (or modeled) variance, Snijders and Bosker (1994, 1999) defined the modeled variance in terms of the proportion reduction in the mean squared prediction error. Snijders and Bosker's measures have been shown to have monotonic properties in the population (i.e., cannot decrease when predictors are added). However, their approach is only applicable to models with a random intercept term, not random slope terms at Level 2.

GROWTH MODELING FROM A MULTILEVEL PERSPECTIVE

When data are collected on multiple occasions over time from individual subjects, multilevel modeling can be used to model longitudinal change. The data can be collected using a fixed measurement schedule for all individuals (e.g., once per day as in daily diary data) or following a unique measurement schedule for each individual (e.g., as in event-contingent data collection). Longitudinal data may contain individual trends or daily or weekly cycles that are not detected if the time-related variable is ignored (West & Hepworth, 1991). In the present case, trends in alcohol consumption over a 10-week period in a normal college student population were not theoretically expected. But trends would be expected in other contexts involving long-term personality or cognitive development (e.g., increases in conscientiousness in adulthood, Jones & Meredith, 1996; decreases in fluid intelligence in older adults, McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002). Trends or changes in level also often occur in contexts in which all participants have been selected to begin the study following a common transition (e.g., college entrance, Tamir et al., 2007) or in anticipation of an upcoming known common major life event (e.g., bar examination, Shrout, Herman, & Bolger, 2006). In contrast, cycles occur when a common daily, weekly, monthly, or other rhythm affects the behavior of the participants. In their original study, Armeli et al. (2000; see their

Figure 1, p. 983) showed that the daily measure of *ALC* systematically changed over the course of the week (highest on Saturday, lowest on Monday). In Model D below, we included a time-related variable (weekday vs. weekend) as a Level 1 predictor to capture the weekly cycle in alcohol consumption.

Model D: Random Coefficient Model With Stress and *WKEND*

We did a preliminary analysis (not reported in Table 2) to see whether alcohol consumption increased on weekends. We created a dummy variable *WKEND* in which "weekend" drinking days (Friday and Saturday) were coded as 1 and other days were coded as 0. Other coding schemes could be used to represent more complex weekly cycles in the data (see West & Hepworth, 1991). Alcohol consumption increases by 0.44 drinks per day during the weekends from the baseline weekday rate of 2.07 drinks per day, so that the mean alcohol consumption was 2.51 drinks per weekend day.

In Model D, we included both *STR* and *WKEND* as Level 1 predictors. Model D allows us to test the relationship between stress and drinking, controlling for the change due to the weekly cycle.

$$\text{Level 1 (within individual): } (ALC)_{ti} = \beta_{0i} + \beta_{1i}(STR)_{ti} + \beta_{2i}(WKEND)_{ti} + e_{ti} \tag{11}$$

$$\text{Level 2 (between individuals): } \beta_{0i} = \gamma_{00} + u_{0i} \tag{12}$$

$$\beta_{1i} = \gamma_{10} + u_{1i} \tag{13}$$

$$\beta_{2i} = \gamma_{20} + u_{2i} \tag{14}$$

In the Level 1 model (equation 11), the time variable *WKEND* is 0 on weekdays, so the intercept β_{0i} represents the predicted alcohol consumption of person *i* who has zero stress on weekdays. The slope β_{1i} represents the relationship of stress to alcohol consumption for person *i*, controlling for the weekend cycle. The slope β_{2i} represents person *i*'s change in the level of alcohol consumption from weekdays to weekends,[6] with $STR = 0$. The Level 2 model (equations 12–14) captures individual differences in β_{0i}, β_{1i}, and β_{2i}. Once again, the

6. If the effect is interpreted as the value when other variables in the equation (here, *STR*) are 0, it is always correct. If there are no higher order terms in the equation as is the case here, it is also true that the effect will be the same regardless of the variable's value. In this case, WKEND can also be interpreted as the effect when *STR* is held constant.

Level 2 residuals, u_{0i}, u_{1i}, and u_{2i}, are assumed to be multivariately normally distributed so that the means, variances, and covariances of the three Level 2 effects fully capture the available information in the Level 1 equation:

$$\begin{pmatrix} u_{0i} \\ u_{1i} \\ u_{2i} \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & & \\ \tau_{10} & \tau_{11} & \\ \tau_{20} & \tau_{21} & \tau_{22} \end{bmatrix} \right).$$

In Model D, we expected that there would be individual differences in the change of alcohol consumption from weekdays to weekends. The residual u_{2i} reflects this expectation. Researchers can choose to specify a simpler model by fixing the variance of u_{2i} to 0 ($\beta_{2i} = \gamma_{20}$). The change of alcohol consumption from weekdays to weekends is now constrained to be constant for every individual (i.e., no individual differences).

Table 2, Model D provides the results and Table 2, Model B provides the appropriate baseline comparison. For Model D, $\hat{\gamma}_{00} = 1.66$ indicates the mean alcohol consumption during the weekdays with zero stress. Alcohol consumption increased significantly during the weekends ($\hat{\gamma}_{20} = 0.43$) with zero stress. Stress was significantly related to drinking even after holding constant the effect of the *WKEND* predictor ($\hat{\gamma}_{10} = 0.21$). Compared to Model B, the Level 1 residual variance was reduced by 15% [(0.40 – 0.34)/(0.40) = 0.15)] by adding *WKEND* as a Level 1 predictor. At Level 2, there were significant individual differences both in the increase in alcohol consumption during the weekends ($\hat{\tau}_{22} = 0.13$) and in the relationship between stress and drinking ($\hat{\tau}_{11} = 0.09$).

OBSERVATIONS ON THE MODELS SO FAR AND THEIR INTERPRETATION

We have four related comments on the models that we have considered so far.

1. The interpretation of the results is critically dependent on the coding of each of the variables and where 0 is located on each predictor.

2. Correspondingly, when the direction, magnitude, and level of statistical significance of a corresponding coefficient are to be compared in two models, care must be taken to be sure all of the other variables are evaluated at the same value in the two models. Often no

attempt to do that is made (as was the case here), so comparisons of the "same" coefficient across models may be confounded (and confusing) because they are comparing different estimates. For example, in Model B $\hat{\gamma}_{10} = 0.25$, $p < .05$, and in Model C $\hat{\gamma}_{10} = -0.03$, *ns*, suggesting different results at first glance. In Model B $\hat{\gamma}_{10}$ is the mean level of alcohol consumption, whereas in Model C $\hat{\gamma}_{10}$ is the intercept, the predicted level of alcohol consumption when $AOE = 0$ and $GEN = 0$ (the average effect combining men and women). Centering predictor variables (see the next section) provides one method of avoiding this interpretational problem.

3. The comparison of Models B and D suggests that the Level 1 variables of *STR* and *WKEND* have only a low positive correlation (here, $r = .08$). In a population in which stress and weekend had a high negative correlation (e.g., lower *STR* on weekends due to reduced workload), the relationship between stress and drinking could be appreciably higher in magnitude if *WKEND* were held constant. On the other hand, in a population in which *STR* and *WKEND* had a high positive correlation (e.g., higher *STR* in individuals without work-related social interactions), the relationship between stress and drinking could be appreciably lower in magnitude. The low positive correlation in the present example also implies that little change from Model C would be obtained when the Level 2 predictors of *GEN* and *AOE* were added to Model D, a result that was found (not reported in Table 2).

4. More complex hypothesized interactions involving the time-related variable (*WKEND*) could be added to the model. For example, if the relationship between *STR* and alcohol consumption were expected to be stronger on weekends because of lessened anticipated consequences of overindulgence, then a $WKEND \times STR$ interaction could be added to the Level 1 model. If a priori hypotheses existed for their inclusion, this Level 1 interaction term could also be permitted to interact with Level 2 predictors.

CENTERING

Centering rescales the predictors to simplify the interpretation of the results. In multiple regression, mean centering is straightforward. For continuous predictors, each predictor X is centered by subtracting its mean from each score, $X_C = X - \bar{X}$, where X_C is the centered predictor (Aiken & West, 1991). For categorical predictors, contrast

codes or effect codes[7] are used (see Cohen et al., 2003, chap. 8). The result is that the value 0 on each predictor now represents the mean value of each variable so that lower order effects in complex equations involving interactions can be interpreted as *average* effects. In our above example, *GEN* was contrast coded, greatly simplifying its interpretation. A full discussion of centering in multiple regression can be found in Aiken and West (1991) and Cohen et al. (2003). West et al. (1996) and West et al. (2007) offered briefer presentations in the context of personality research.

In MLM, centering is both more important and more complex than in multiple regression. It is more important because researchers in MLM often wish to interpret the intercept, particularly in the Level 2 equations. It is more complex because in MLM there are far more centering options, the choice of which can potentially be very consequential for the interpretation of the results. Below we present six brief observations on centering in MLM in common personality designs, but we encourage researchers to study the more complete accounts presented in the references cited below.

1. For predictors having values of 0 that represent the absence of the variable ($STR = 0$ means there is zero stress), interpretations at a value of 0 will often be meaningful. In contrast, when 0 is an arbitrary number on the predictor (e.g., 0 on a Likert scale where 1 is *disagree very much* and 6 is *agree very much*), the results are not interpretable.

2. Mean centering can be useful in model comparison, as the interpretation of the terms does not change. Our earlier cautions about the meaning of key terms in comparing the results of Models A and B and Models B and D would be unnecessary if terms in both models had been mean centered.

3. When longitudinal growth models are employed, 0 on the time variable is usually coded to represent a meaningful time point in the study. For example, in the Tamir et al. (2007) study of adaptation to college, 0 might be chosen to represent the start of the semester or 0

7. Unweighted effect codes are centered at the unweighted mean $\bar{X}_U$ of the groups $\bar{X}_U = (\bar{X}_{G1} + \bar{X}_{G2} + \bar{X}_{G3})/3$, where $\bar{X}_{G1}$, $\bar{X}_{G2}$, and $\bar{X}_{G3}$, respectively, are the means of the three groups. Weighted effect codes are centered at the weighted mean of the groups $\bar{X}_W = (n_{G1}\bar{X}_{G1} + n_{G2}\bar{X}_{G2} + n_{G3})/(n_{G1} + n_{G2} + n_{G3})$, where n_{G1}, n_{G2}, and n_{G3}, respectively, are the sample sizes of the three groups. West, Aiken, and Krull (1996) provided a full presentation of weighted and unweighted effect codes and their interpretation.

might be chosen to represent the end of the semester. Biesanz, Deeb-Sossa, Papadakis, Bollen, and Curran (2004) offer advice on centering in growth models.

4. In a two-level model, if centering of Level 2 variables is being considered, standard mean centering is always preferred. Such centering can facilitate interpretation of the Level 2 effects and cross-level interactions (Enders & Tofighi, 2007).

5. Centering at Level 1 involves a choice between centering options that must be made on the basis of the multilevel design, the question being addressed, and the theoretical conception in the area. The basic choice is between analyzing the raw data (as we did above), centering at the grand mean, and centering within context (within-person centering). Centering at the grand mean parallels mean centering in regression. Raw score and grand mean centered analyses might be chosen if the researcher is focused on the effects of the absolute measured values of stress on alcohol consumption. Here, a value on the *STR* variable of 3 relative to a stress value of 1 would be expected to have the same impact on drinking regardless of the person's usual (mean) level of stress. In contrast, centering within context (person) involves computing the mean of each person's 70 days of stress data and then calculating the deviation of each day's value on the *STR* variable for person *i* from his or her usual (mean) level. In our example, centering within person tests a hypothesis that deviations from the person's *usual* level of stress affects drinking. Centering within person will often be the best choice when the primary focus is on the effect of the Level 1 variables (Enders & Tofighi, 2007). However, as Kreft, de Leeuw, and Aiken (1995) noted, simple centering within person removes all of the between-persons variation from the data—information on each person's mean (chronic) level of stress is lost. Kreft et al. (1995) described a second version of centering within person (CWC2) that uses each person's mean value of stress as a predictor at Level 2. The result of this second version of the centering-within-person procedure is to restore this information. An important result is that the effect of stress on drinking is now neatly partitioned into two parts: (a) the effect of daily deviations from each person's mean level of stress on drinking (within-person component) and (b) the effect of each person's mean (chronic) level of stress on drinking (between-persons component representing the effect of stable individual differences). These two parts correspond to the two classic lines of personality research: person-based research

and individual differences. We recommend that personality researchers use this second version of centering within person (context, CWC2) unless they can claim that individual differences in mean level on the predictor are of absolutely no theoretical interest.[8]

6. Centering within context may be based on unreliable estimates of the cluster (person) means, particularly when the number of observations per cluster and the ICC are both small. An alternative multilevel latent variable approach may be preferred under some conditions because it can lead to less biased estimates of person-based (contextual) effects (see Lüdtke et al., 2008).

MODEL E: STRESS CENTERED WITHIN PERSON

To illustrate the CWC2 analysis, we performed the CWC2 analysis corresponding to Model C presented earlier. At Level 1, stress centered within person (STR_CWC) was the predictor. At Level 2, GEN, AOE, $GEN \times AOE$, and centered mean person stress levels ($\overline{STR}_i$) were the predictors. Our equations for Model E are

$$\text{Level 1 (within individual): } (ALC)_{ti} = \beta_{0i} + \beta_{1i}(STR\_CWC)_{ti} + e_{ti} \tag{15}$$

Level 2 (between individuals):

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(GEN)_i + \gamma_{02}(AOE\_C)_i + \gamma_{03}(GEN)_i(AOE\_C)_i + \gamma_{04}(\overline{STR}_i) + u_{0i} \tag{16}$$

$$\beta_{1i} = \gamma_{10} + \gamma_{11}(GEN)_i + \gamma_{12}(AOE\_C)_i + \gamma_{13}(GEN)_i(AOE\_C)_i + \gamma_{14}(\overline{STR}_i) + u_{1i} \tag{17}$$

where $\overline{STR}_i$ is the person mean of stress for person i centered at the grand mean, STR_CWC is stress centered at the person mean

8. As described above, analyzing the raw data, centering around the grand mean, and centering within context produce different solutions in terms of their interpretations. However, the information about the fixed effects represented in these three approaches (raw, grand mean centered, CWC2) is algebraically equivalent, so that the fixed-effects part of the solution of one approach can be transformed algebraically.

within each individual, and *AOE_C* is the *AOE* centered at the grand mean. The results are reported in Table 2, Model E.

There are two key differences between Model C and Model E that have dramatic effects on the interpretation of the results. First, all Level 2 predictors have been mean centered so coefficients are interpreted at the mean value of each of the Level 2 variables. The effects now represent average effects of the predictor. Second, daily stress levels have been centered within person at Level 1 and the mean stress level for each person is added as a predictor at Level 2. This partitions the effect of stress into an average within-person effect (how do deviations from the person's average level of stress affect alcohol consumption?) and a between-persons individual differences effect (how does each person's average level of stress affect alcohol consumption?). These two components can differ dramatically in some data sets.

In Table 2, Model E, we initially focus on effects related to the intercept, which represents each person's mean level of alcohol consumption in this model. The estimated grand intercept $(\hat{\gamma}_{00}) = 2.19$ indicates the predicted mean alcohol consumption in the sample. Given that each participant's mean was based on the same number of observations (70 days), this value equals the intercept from Model A. The negative coefficient for *GEN*, $\hat{\gamma}_{01} = -0.87$, indicates that the mean level of alcohol consumption was 0.87 drinks higher for men than women, given men were coded -0.5 and women were coded $+0.5$. *AOE* was positively related to *ALC*, $\hat{\gamma}_{02} = 0.49$, $p < .05$, indicating individuals with higher *AOE* had higher individual mean levels of alcohol consumption across the 70 days of the study. These *GEN* and *AOE* average effects were modified by a *GEN* × *AOE* interaction, $\hat{\gamma}_{03} = -0.38$, indicating that the relationship between *AOE* and *ALC* was stronger for men than women. Finally, each person's mean level of stress was *negatively* associated with mean alcohol consumption, $\gamma = -0.82$, $p < .05$, when *AOE* was at its mean value and *GEN* was ignored. Figure 2 shows the estimated relationship of *STR_CWC* and *ALC* at three different values of individual mean stress, the individual's typical or "chronic" stress level. $STR\_CWC = 0$ indicates each individual's mean stress level because *STR_CWC* was centered at each individual's mean. When $STR\_CWC = 0$, the predicted alcohol consumption is highest when person mean (chronic) stress was at (Mean $-$ 1 *SD*) and lowest when person mean stress was at (Mean $+1$ *SD*).

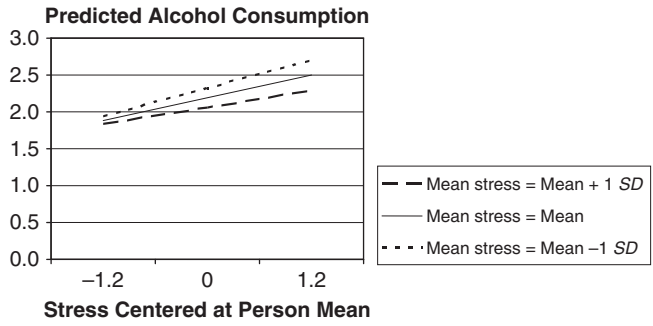

Figure 2
Plot of interaction between within-person and between-persons components of stress on alcohol consumption. The *x*-axis portrays daily deviations from each person's own mean level of stress. The three lines represent different mean (chronic) stress levels (between-persons differences).

We now turn to the effects of positive deviations (greater than usual stress for person *i*) from that individual's mean level of stress on alcohol consumption. The larger the positive deviation (more stress), the greater alcohol consumption was on that day on average, $\hat{\gamma}_{10} = 0.25$, $p < .05$. This effect was greater for men than women, $\hat{\gamma}_{11} = -0.32$, $p < .05$, indicating a steeper slope for men relative to women for the effect of unusually high levels of daily stress on drinking during that day. Higher values of *AOE* were also associated, $\hat{\gamma}_{11} = 0.12$, with a steeper slope (stronger relationship) between positive deviations from the usual level of daily stress and alcohol consumption that day. An interaction between *GEN* and *AOE* indicated that, for the relationship between *AOE* and higher than usual values of stress in producing drinking that day, the effect was stronger for men than women. Finally, there was an interaction between each individual's mean level of stress and positive deviations from his or her mean level of stress on drinking that day, $\gamma = -0.41$. As depicted in Figure 2, as the person's mean (chronic) level of stress increased, the effect of higher than usual levels of stress that day on drinking that day became less. These results illustrate how the effects of one's mean level of stress (between individuals) and deviations from one's own average level of stress (within individual) can be partitioned into between-individuals and within-individual effects, sometimes showing different results.

ESTIMATION: ISSUES AND CONCERNS

In multiple regression, the use of ordinary least squares (OLS) to estimate parameters yields a unique algebraic solution. The estimates of parameters and their standard errors are appropriate for small sample sizes. The assumptions of OLS focus on the residuals (e.g., homoskedasticity). In contrast, MLM uses maximum-likelihood estimation (see Enders, 2005, for a readable introduction). Maximum likelihood does not yield a simple algebraic solution but instead requires the use of a numerical algorithm to search for a solution. Maximum likelihood is based on large sample statistical theory; the performance of maximum likelihood may deteriorate in smaller samples so that the estimates of parameters or standard errors in MLM may no longer be accurate. Finally, maximum likelihood makes strong distributional assumptions involving multivariate normal distributions. The benefit of these differences between the two estimation methods is that maximum-likelihood approaches permit the estimation of more complicated and often more realistic models than OLS. The cost of these differences is that very careful monitoring of the results of maximum-likelihood analyses is needed. Three implications are given below.

1. The standard test of parameters reported by MLM programs is the Wald test, which is an approximate large sample test. More accurate tests can be achieved by model comparison procedures (likelihood ratio tests) described in the next section.

2. The numeric search procedures used in maximum likelihood do not always find a solution. In general, these procedures start with "initial guesses" (start values) of what the values of the parameters in the solution might be. Sophisticated numeric procedures are then used to find better values that increase the likelihood that these parameter values are consistent with the observed data. At some point, further improvement is not possible: When that point is reached, estimation is said to have converged and this final solution is reported. However, if the model is not correctly specified, the sample size is too small, the data do not follow the assumed (multivariate normal) distribution, or the numeric search procedure is given poor start values, problems may occur. In such cases, the model may fail to converge and no solution will be found. Modern computer programs typically include "fixes" that permit estimation when problems arise; these fixes result in a warning message. Some warning

messages indicate the solution is untrustworthy. Other messages indicate that the computer has taken a reasonable action (e.g., setting the variance of the slope to 0 when there does not appear to be any slope variance) that permits estimation typically without appreciable effects on the results. All parameter estimates and standard errors, even ones of little theoretical interest, should be examined. Unreasonable estimates of parameters, very large standard errors, or large numbers of iterations (numeric steps required to find a solution) may indicate problems.

3. In MLM, problems of estimation increase with small sample size, unbalanced designs, and missing data. In two-level MLM, the sample size at each level needs to be considered. At Level 2, a minimum of 20 Level 2 units is a rule of thumb recommendation for proper estimation[9] (e.g., Kreft & de Leeuw, 1998). This minimum number will depend on the model, the data structure, and the computer software. Of course, more Level 2 units will always be better. This minimum recommendation presents little difficulty in personality research designs in which persons are the Level 2 units, but it can be a practical limitation in designs in which the clusters are classrooms, schools, or friendship groups. Note that the statistical power to test Level 2 effects follows the number of Level 2 units, so that a much larger number of Level 2 units than 20 will typically be needed to test key hypotheses at Level 2 with adequate statistical power (e.g., stable individual differences, variance components). At Level 1, very small numbers of observations per Level 2 unit may be used in some designs. As the number of observations per unit decreases, the estimates of the parameters at both Level 1 and Level 2 are not problematic, but the estimates of the variances and covariances and their standard errors of the Level 2 effects can become biased (Hox & Maas, 2001; Ryu, 2004). This issue is not a problem in areas such as daily diary studies in which 20 or more observations per individual are collected, but it may lead to appreciable bias in tests of random effects in areas such as studies of longitudinal

9. When fewer than 20 units are available at Level 2, Kreft and de Leeuw (1998) recommended using dummy codes to represent the clusters. The dummy variable procedure addresses intercept differences at Level 2 but does not address slope differences.

growth in which only 3 or 4 observations are collected or in studies of dyads in which by definition only two people compose each cluster. These problems become more serious if there are a different number of persons in each Level 2 unit or if data are missing, resulting in an unbalanced design. The estimates of an MLM based on 400 total observations will be far better if they are (a) based on 100 participants each with 4 observations each rather than (b)100 participants with a mean of 4 observations each, but varying from 1 to 8 observations per person (see Mok, 1995). A very limited number of observations at Level 1 may severely restrict the complexity of hypotheses that may be tested. Sometimes, as in the study of dyads, restrictions may need to be placed on the model to permit estimation of even basic relationships. The substantive reasonableness of each restriction needs to be considered, as the results of the entire model may depend critically on the specified restriction. Finally, the statistical power of the design to test hypotheses at Level 1 will be related to both the total number of observations and the degree of clustering. When the design is balanced so that each Level 2 unit has an equal number of Level 1 observations, the effective N ($N_{\text{effective}}$) is defined as

$$N_{\text{effective}} = \frac{n_{\text{L1}} n_{\text{L2}}}{(1 + (n_{\text{L1}} - 1)ICC}, \tag{18}$$

where n_{L1} is the number of Level 1 observations per Level 2 unit, n_{L2} is the number of Level 2 units, and ICC is the intraclass correlation coefficient that assesses the degree of dependency. When $ICC = 0$, indicating no clustering, $N_{\text{effective}} = n_{\text{L1}} n_{\text{L2}}$, the total number of observations in the study. As the ICC increases in value, $N_{\text{effective}}$ decreases, becoming progressively lower than $n_{\text{L1}} n_{\text{L2}}$, eventually approaching n_{L2} as the ICC approaches its maximum value of 1.0. $N_{\text{effective}}$ can be used to estimate the statistical power of the proposed multilevel design. Power analysis programs for MLM, such as the freestanding PINT program (Bosker, Snijders, & Guldemond, 2003) and MLPowSim routine within the MLwiN program (Browne, Golalizadeh-Lahi, & Parker, 2009), permit researchers to calculate statistical power and to investigate the trade-offs between having different numbers of Level 1 (e.g., number of days) and Level 2 (e.g., number of persons) units in planning their studies. Monte Carlo studies of statistical power can also be conducted in MPlus (see Muthén & Muthén, 2002).

MODEL COMPARISON PROCEDURES

Model comparison procedures are used when the researcher wishes to compare a more complex model with a simpler model in which one *or more* of the parameters have been set to 0. For example, Model D, which estimated Level 1 effects of *STR* and *WKEND*, could be compared with Model B, in which *WKEND* was not included in the model ($\beta_{2i} = 0$). Model B is described as being *nested* within Model D. Two nested models can be accurately compared using the likelihood ratio test.

The maximum-likelihood estimation procedure produces a measure known as the deviance, which is a measure of lack of fit of the model to the data. The difference between the deviances of the two models provides an accurate test of the parameters that are set to 0. The likelihood ratio test (LR) is

$$\text{LR} = \textit{Deviance}_1 - \textit{Deviance}_2, \tag{19}$$

where $\textit{Deviance}_1$ is the deviance of the model with fewer estimated parameters and $\textit{Deviance}_2$ is the deviance of the model with more estimated parameters. LR is evaluated against a χ^2 distribution with *df* equal to the number of predictors added to the model. For the test of the *WKEND* effect, the deviance for Model B was 13,238.7 and the deviance for Model D was 13,039.9, so LR = 198.8, $df = 1$, $p < .001$. The LR test also permits a test of the addition of a set (more than 1) of predictors to the model.

Several qualifications of the LR test are needed when tests of variance components are involved. First, if only variance components are involved, an alternative estimation method known as restricted maximum likelihood (REML) will be slightly more accurate than standard maximum likelihood. Second, tests of variance components should always be one-tailed; variances *cannot* take on negative values. Two-tailed *p* values reported by computer programs should be divided by 2. Third, the LR test does *not* follow a chi-square distribution when the parameter being tested is set at a boundary value. The smallest possible value of a variance is 0, so tests that a variance component is 0 represent a commonly encountered boundary condition. Tests of covariances or fixed effects that can potentially take on positive, zero, or negative values do not share this problem. Stoel, Garre, Dolan, and van der Wittenboer (2006)

presented an exact method for testing parameters that are set to boundary values.

The LR test cannot be used to directly compare models that are not nested. Instead, two different model comparison indices known as Akaike's Information Criterion (AIC; Akaike, 1973) and the Bayesian Information Criterion (BIC; Schwarz, 1978) are commonly used. Both indices combine information from the deviance and a penalty function based on the number of parameters in such a way that more parsimonious models are favored. In other words, if two models have similar deviances but one requires fewer model parameters, the model with fewer parameters will be preferred. The expressions for the two model comparison indices are as follows:

$$\text{AIC} = \textit{deviance} + 2q \tag{20}$$

$$\text{BIC} = \textit{deviance} + q\ln(n) \tag{21}$$

where q is the number of parameters being estimated and n is the sample size. As q increases, indicating a less parsimonious model, AIC and BIC will increase. In general, the BIC has a stronger penalty function than the AIC except at very large sample sizes, so it tends to prefer less complex models. The AIC tends to be easier to use than the BIC in the MLM context because it does not involve sample size.[10] Recall that the sample sizes at Level 1 and Level 2 differ, making some comparisons using the BIC more difficult. In practice, model comparison indices are often reported for a set of models, and the model with the smallest relative AIC and BIC is chosen. In our example, Model C had predictors *STR* at Level 1 and *GEN* and *AOE* at Level 2. Model D had predictors *STR* and *WKEND* at Level 1 and no predictors at Level 2. Comparing these two non-nested models: Model C, AIC = 13,983.7, BIC = 13,994.1; Model D, AIC = 12,979.4, BIC = 13,011.6. Given the smaller values of both the AIC and BIC, Model D would be preferred on the grounds of providing a more parsimonious explanation of the results.

10. For BIC, SPSS uses the total number of observations ($n_{L1}n_{L2}$); SAS uses the number of Level 2 cases (n_{L2}).

ASSUMPTION TESTING IN MULTILEVEL MODELS

Violations of assumptions provide a signal that there may be problems with the model being tested, the data, or both. When serious violations of assumptions are detected, researchers need to probe potential reasons for the violations. Below we offer some simple graphical methods of detecting such violations. Graphical methods have the advantage of being able to detect a wide variety of violations, but they do not yield significance tests. The magnitude of the violation may be more important than its statistical significance. Modest violations of the assumptions of homoskedasticity and normality of the Level 1 and Level 2 residuals in MLM do not greatly affect parameter estimates or their standard errors (Raudenbush & Bryk, 2002).

Probing the Homogeneity Assumptions

MLM assumes that the variances of the residuals are constant around the regression lines at both Levels 1 and 2. At Level 2, the standardized residual dispersion[11] d_j for each cluster may be calculated (Raudenbush & Bryk, 2002):

$$d_j = \frac{\ln(S_j^2) - \left[\sum f_j \ln(S_j^2) / \sum f_j\right]}{\sqrt{(2/f_j)}}, \tag{22}$$

where S_j is the estimated residual standard deviation of cluster j, and f_j is the degrees of freedom associated with S_j. Two complementary graphical techniques are useful in visualizing major violations of homoskedasticity. Raudenbush and Bryk (2002) suggested a q-q (quantile-quantile) plot that displays the expected value of the dispersion measures on the y-axis against the standardized residual dispersion on the x-axis for all of the Level 2 clusters in the sample (see Cohen et al., 2003, pp. 137–141, for a discussion of q-q plots). If the homogeneity assumption were perfectly met, there would be a 45-degree straight line between the expected and the standardized dispersions. Heteroskedasticity is indicated when all or a substantial

11. The HLM program (Raudenbush, Bryk, Cheong, Congdon, & du Toit, 2004) offers the overall test of the homogeneity of variance which uses d_j. HLM does not directly produce the individual d_j. Hoffman (2007) provided SAS syntax (http://psych.unl.edu/hoffman/homepage.htm) that can be used for obtaining individual d_j and testing the homogeneity assumption (Snijders & Bosker, 1999).

number of points follow a curved line. A second plot is useful in identifying outlying clusters with high (or low) variance. A histogram of the natural logarithm of the standardized residual dispersions is plotted. Outlying clusters will be distinct from the remainder of the distribution. Note that SPSS constructs what it terms a "p-p" plot. The *x*- and *y*-axes of the plot must be reversed to construct a proper q-q plot (see the Web site).

One important source of heteroskedasticity in the Level 2 residuals is that the model may not be correctly specified at Level 1. Some examples of misspecification at Level 1 are as follows: (a) an important Level 1 predictor is not included in the model, (b) there are subgroups (e.g., unmodeled important differences in *STR* between social and nonsocial situations in Level 1 data), or (c) the Level 1 error structure is not correctly specified. Problem (c) could occur in diary data if major relative to minor events were associated with more response variability or if responses that occurred closer in time were more similar (e.g., serial dependency; see West & Hepworth, 1991).

These possibilities can be probed using the overall model deviance statistics to compare a model with homogeneous Level 1 residual variance to models with potentially heterogeneous variances. In empirical studies using standard MLM designs in which participants are nested within larger groups, Hoffman (2007) examined possibility (a), adding the Level 1 variable of mental functioning to her model. Raudenbush et al. (2004) explored possibility (b), using gender subgroups at Level 1. In the context of growth curve modeling, Kwok, West, and Green (2007) explored possibility (c), showing how dependency in the error structure can affect tests of the growth parameters.

Probing the Normality Assumptions

As noted earlier, assumptions are made about the normality of the random effects at Level 2. Two methods are commonly used to identify major violations of these assumptions. The Mahalanobis distance is a multivariate measure of the distance of a point from the centroid, the point at which each of the variables has its mean value (M_1, M_2, . . ., M_p). In the present context, the Mahalanobis distance represents the multivariate distance of a point representing the estimated intercept and slope for case *i* from the mean intercept and slope. The Mahalanobis distances for each of the cases from the

centroid at Level 2 (here, persons) are ordered from lowest to highest. A q-q plot is constructed in which the actual Mahalanobis distances for the estimated Level 2 random effects are on the *y*-axis and the ordered random effects *if* the data had a multivariate normal distribution are on the *x*-axis. If the normality assumption were perfectly met, there would be a 45-degree straight line representing the relationship between the expected and the estimated Mahalanobis distances. Once again, if the points corresponding to each cluster fall on a curved line or if several of the Level 2 clusters are located far from the 45-degree line, the normality assumption has been violated. Figure 2 displays the q-q plot for the alcohol example in Model E (described in an earlier section). As can be seen, the points for each person (Level 2) fall close to the 45-degree line ($R^2 = .959$; see the Web site).

An alternate approach is to probe the normality assumption for the random effects by examining the discrepancy between the maximum likelihood and the robust estimated standard errors (*SE*s). When the normality assumption is met, the maximum

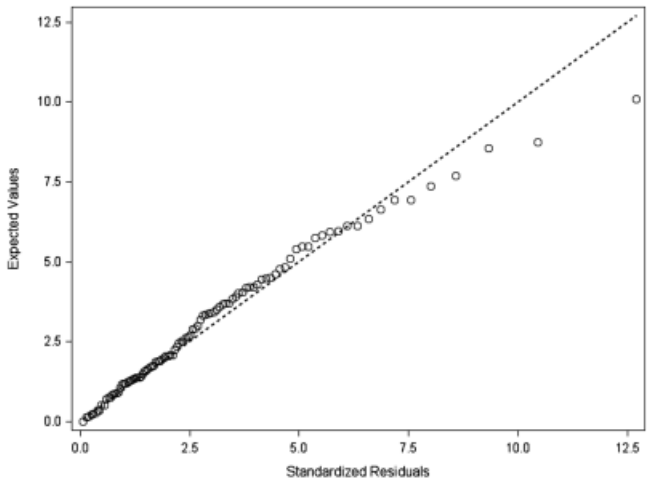

Figure 3
Q-q plot of Level 2 residuals against a normal distribution. If the distribution is exactly normal, the residuals will exactly fit the 45-degree straight line. The plot illustrates only minor deviations from normality.

likelihood and the robust standard errors are expected to be very close to each other. On the other hand, the normality assumption may be violated if substantial discrepancy occurs between the robust and the nonrobust *SE*s. We compared the parameter estimates and standard errors between the maximum likelihood and the robust results and found no appreciable differences between the results, again indicating no major discrepancies from normality in our example. Taken together, our probing for violations of the homoskedasticity and normality assumptions did not turn up any evidence of model misspecification or problems with the data in our example.

DIAGNOSTIC MEASURES IN MULTILEVEL MODELS

As in single-level multiple regression, outliers have the potential of greatly influencing the results of analyses (see Cohen et al., 2003, chap. 10, for a discussion of diagnostic measures in multiple regression). However, because of the multilevel structure, the effect of errant data points in MLM is far less clear. In two-level MLM, the errant cases can come from extreme observations within clusters, entire clusters may be extreme, or both. The new MLM diagnostic measures are used in a manner paralleling that of the analogous diagnostic measures in OLS regression. Shi and Chen (2009) present a fuller description, including the formulas for each of the MLM measures and the corresponding diagnostic measures in OLS regression. Several of the new measures have been recently developed so that they are only now beginning to be implemented in some of the standard statistical packages.

Most authors (e.g., Lewis & Langford, 2001; Snijders & Berkhof, 2008) recommend a top-down diagnostic procedure that moves from the highest to the lowest level in MLM. There are two general steps: (a) The potential extreme cases are identified by the diagnostic measures, and (b) dummy variables are created to differentiate the extreme cases from other cases. The dummy variables are added to the original model and tested for statistical significance. A case is confirmed as an errant point if the corresponding dummy variable is significant. These procedures are repeated at each level to identify the errant cases. Space limitations permit us to describe these procedures only at Level 2.

Leverage

In OLS regression, leverage is used to assess how far a particular data point is from the centroid of the predictors. Leverage at Level 2 in MLM can be calculated based on complex expression involving the matrix containing the values of all of the predictors for each participant and the matrix containing the variances and covariances of all of the random effects (Langford & Lewis, 1998). Clusters with leverage values larger than a cut-off value of $2.5/J$ are identified as potential outliers, where J is the number of Level 2 clusters. Based on Model E, we examined the leverages of our alcohol use example and found that all the leverages were smaller than the suggested cut-off value.

Externally Studentized Residuals (ESR)

Paralleling externally studentized (deletion) residuals in the OLS regression (see Cohen et al., 2003, chap. 10), ESR in MLM uses deletion residuals (i.e., $d_{qj} = y_j - \hat{y}_{j(j)}$). Otherwise stated, the predicted value ($\hat{y}_{j(j)}$) is based on the fitted model excluding the jth cluster (the (j) notation indicates this cluster is excluded). Excluding the jth cluster prevents that cluster (possibly an extreme case) from influencing the intercept or slope of the regression line, resulting in a purer measure of its extremity. This deletion residual is then divided by the adjusted standard error ($s_{qj(j)}$), which is also based on the model without the jth cluster:

$$ESR_{qj(j)} = \frac{d_{qj}}{s_{qj(j)}} \tag{23}$$

The distribution of ESR is approximately normal for large samples. Depending on the number of clusters, a cut-off value of $ESR_{qj(j)}$ can be chosen to represent an extreme value. For typical MLM designs with 100 or fewer Level 2 units, Lewis and Langford (2001) suggested that cases with an absolute value of $ESR_{qj(j)}$ greater than 3.0 can be classified as outliers. As of this writing, the ESR diagnostics at both Level 1 and Level 2 are fully implemented in the MLwiN program and are available only at Level 1 in SAS.

DFFITS

A final commonly used measure in OLS regression is *DFFITS* (Belsley, Kuh, & Welsch, 1980; see also Cohen et al., 2003, chap. 10),

which provides a global measure of how much each case influences the results of the regression equation. *DFFITS* compares the predicted value for case (i) with case (i) included versus not included in the data set. A MLM version of *DFFITS* is available at Level 2 to assess the extent to which cluster *j* influences the outcome. *DFFITS* is calculated by the following equation:

$$DFFITS_j = \frac{\hat{y}_j - \hat{y}_{j(j)}}{s_{j(j)}\sqrt{h_{jj}}} \tag{24}$$

Here the deletion residual for cluster *j* is in the numerator, and the denominator contains the adjusted standard error $s_{qj(j)}$ and the square root of the leverage h_{jj} for cluster *j*. One plausible cut-off value for *DFFITS* is $2\sqrt{\frac{q+1}{J}}$, where q = number of Level 2 predictors and J is the number of Level 2 clusters in the model. Cases with absolute values of *DFFITS* larger than the recommended cut-offs are considered as influential points. In the alcohol use example, again using Model E, none of the Level 2 *DFFITS* values exceeded this cut-off value. Another similar measure of detecting influential cases in OLS regression is Cook's distance, which has been extended to two-level MLM by Lesaffre and Verbeke (1998) and Verbeke and Molenberghs (2000).

Addressing Outliers: Some Possible Remedies

Outliers can potentially seriously alter the results of an MLM analysis. When an outlier is detected, the researcher should take steps to attempt to identify its source. In some cases, outliers may represent errors in the data set. These should be deleted or corrected. In other cases, outliers may represent participants from a different population (e.g., an intellectually gifted 12-year-old student in a study of sexual attitudes of college students). These cases should be deleted and the inclusion criteria for the population of interest more clearly defined (e.g., 18–22-year-old college students). In other cases, the source of the outlier will be unknown despite careful detective work. Particularly when measures of influence indicate that the outlier is seriously affecting the results of the analyses, remedies should be pursued. Two possible remedies include (a) running the analysis with and without the outlier and reporting both results and (b) the use of robust estimation procedures that are less sensitive to outliers. Whatever remedy is taken, the presence of the outlier should

be clearly reported in the article. Procedures for addressing outliers in the single-level regression case can be found in Cohen et al. (2003, chap. 10).

ADVANCED DEVELOPMENTS: ADDRESSING COMPLEXITIES AND NEW RESEARCH QUESTIONS

In this section, we highlight several newer developments in multilevel modeling. Each of these permit personality researchers to address complex forms of data and specific new research questions that may be of interest.

Three-Level Models

We have considered two-level models in detail through this article. Multilevel modeling can also be extended to cases in which the number of levels in the hierarchy is three or more. A three-level data structure can occur in various ways. To cite three examples, ratings from multiple peers (Level 1) are collected from students (Level 2) nested within classes (Level 3); diary reports are collected on a daily basis (Level 1) from adolescent children (Level 2) nested within families (Level 3); or survey data are collected from individuals (Level 1) within households (Level 2) within neighborhoods (Level 3). With three levels, the random intercept model (a.k.a. the unconditional model) partitions the variance into three components: Level 1, Level 2, and Level 3 variance. The intraclass correlation (ICC) can be computed for Level 2 (proportion of Level 2 variance to the total variance) and for Level 3 (proportion of Level 3 variance to the total variance). Predictors may be added to the model at each level, extending our earlier presentation for the two-level model. In our first example of peer ratings above, length of acquaintance of the peer with the target individual could be a Level 1 predictor, the extroversion of the target individual could be a Level 2 predictor, and type of class (e.g., math, English, gym) might be a Level 3 predictor. Raudenbush and Bryk (2002, chap. 8) provided a more detailed formulation and illustration of three-level multilevel modeling.

Cross-Classified Random Effects Model (CCREM)

Multilevel data do not always follow a strictly hierarchical structure (Rasbash & Browne, 2008; Raudenbush & Bryk, 2002). In the classic

three-level structure, students are clearly nested within small neighborhoods, which in turn are nested within school districts. But students from the same neighborhood may go to different schools, and not everyone from the same school may reside in the same neighborhood—schools and neighborhoods represent crossed rather than nested factors in this data structure. This type of data structure is termed a cross-classified structure.

To analyze this type of data properly, one needs to use the cross-classified random effects model (CCREM; Goldstein, 1986, 1995; Rasbash & Goldstein, 1994; Raudenbush, 1993). Beretvas (2008) provides a brief overview of CCREM along with syntax for analyzing CCREM in several statistical packages, including HLM, SAS, and SPSS. In general, the random effects of the crossed factors (in our example, the neighborhood and school factors) are included in the model so that the corresponding variances of the crossed factors can be properly estimated, and the estimation of the fixed-effect parameters and the corresponding standard errors will be unbiased.

Two major issues may occur in CCREM. First, some of the conditions created by crossing the two factors may have very few, if any, cases (data sparseness). Ninety percent of the children from a specific neighborhood may go to one school, with small numbers (if any) of children attending other schools. Such extreme multicollinearity of the neighborhood and school factors can make it difficult to estimate the model, difficult to interpret the effects, and difficult to achieve sufficient statistical power to detect actual effects when they do exist. Second, it may be difficult to obtain the full information identifying each student's level on each factor. School information may be easily obtainable, but neighborhood information may be more difficult to ascertain, leading researchers to leave it out of the model. Luo and Kwok (2009) and Meyers and Beretvas (2006) have pointed out that the omission of an important classification factor in the design can result in the variance of the ignored factor being redistributed to other levels (student level and school level). This problem can result in the overestimation of variance components at the other levels, greatly reducing the statistical power of tests of fixed effects from the remaining data levels. On the other hand, if a predictor from the ignored data level (e.g., neighborhood) is included in the analysis, the standard errors of the fixed effects of this predictor will be underestimated, leading to Type 1 errors and incorrect statistical inferences. Such findings highlight the potential importance of

collecting cluster information and properly analyzing nonhierarchical data structures.

Nonstandard (Limited) Dependent Variables: Dichotomies, Ordered Categories, and Counts

We noted earlier that MLM makes the assumption that the Level 1 and Level 2 residuals are normally distributed and homoskedastic (constant variance). With measures that approximate continuous variables, modest violations of these assumptions have little impact on the analysis. However, as in multiple regression, a number of commonly used types of dependent variables can yield serious violations of the assumptions. These dependent variables are not continuous but have an upper, lower, or both boundaries so they cannot have a *potential* range from $-\infty$ to $+\infty$. They have a variance structure that is nonconstant or that is severely non-normal. These problematic outcomes are known as limited (a.k.a. "funny") dependent variables; they can be analyzed using the generalized linear model (see Cohen et al., 2003, chap. 13; Coxe, West, & Aiken, in press, for introductions). Three common examples are presented below.

1. Dichotomous outcomes (normal vs. clinical case; fail vs. pass) can only take on observed values of 0 or 1; predicted values can only fall in the range between 0 and 1; the variance of the residuals follows the binomial distribution. Predicted values near 0 or 1 have small variances, whereas predicted values near .5 have relatively large variances.

2. Ordered categories (e.g., a 3-point rating scale for children ranging from 0 [*none*] to 1 [*a little*] to 2 [*a lot*], or a single 5-point Likert item) can only take a small number of integer values, predicted values must fall in a small range, the scale is typically not believed to approximate equal intervals, and the variance might not be constant.

3. Counts (e.g., number of aggressive acts observed in a free play period; number of marriages in lifetime) can take only the values of 0, 1, 2, . . . (0 and positive integers), predicted values can fall only in the range 0 to $+\infty$, and the variance typically increases with the predicted value of the count. The smaller the mean predicted count, the more likely that the standard MLM approach can produce seriously misleading results. A mean count of 10 is often taken as a rule

of thumb minimum value for good performance of standard regression (Coxe, West, & Aiken, in press).

With limited dependent variables, more satisfactory results can be produced by using multilevel generalized linear models (Raudenbush & Bryk, 2002; Skrondal & Rabe-Hesketh, 2004). Multilevel generalized linear models build on the basic MLM approach, except that two features are added that address the problematic features of the data at Level 1.

1. The prediction equation at the Level 1 model looks exactly like the standard MLM model, except that a different outcome variable is used. The outcome variable is a transformation of the original outcome that yields linear relationships, termed a link function. For dichotomous data, the logit $= \log\left(\frac{\hat{Y}_d}{1-\hat{Y}_d}\right)$ is used as the outcome, where $\hat{Y}_d$ is the predicted odds that the dichotomous outcome = 1 (case) given the observed values of the predictors. For counts, the $\ln(\hat{Y}_c)$ is used as the outcome, where $\hat{Y}_C$ is the predicted value of the count variable.

2. A variance function is added to the model that represents the structure of the variance with the type of data under consideration. For dichotomous outcome variables, a binomial function is typically used. The variance is related to the predicted probability that the observation is a case $\hat{P}(Y = 1)$, $Variance = [\hat{P}(Y = 1)][1 - \hat{P}(Y = 1)]$, which yields small variances when the predicted probability is near 0 or 1 and large variances when the predicted probability is near 0.5. For count outcome variables, the variance is related to $\hat{Y}_C$, so that the variance of the residuals increases as the predicted count increases.

With these two additions to the Level 1 model, appropriate statistical tests of all Level 1 and Level 2 effects can be conducted. Such analyses at present can only be conducted in the multilevel modeling context using specialized software, including GLLAMM, HLM, MPlus, and R.

Modeling Nonlinear Growth

Applications of growth modeling in mainstream personality research have focused on linear growth. At Level 1, the basic linear growth model is

$$Y_{ti} = \beta_{0i} + \beta_{1i}(Time)_{ti} + e_{ti} \tag{25}$$

The *i* subscript on *Time* is dropped in many common designs in which each person is measured at the same fixed set of time points

(e.g., daily dairy studies with complete data). Yet, over more extended time periods, outcomes that continue to increase or decline in a linear manner are rare. Height may increase approximately linearly over a short period of childhood, but not over a lifetime. Consequently, models to represent more complex forms of growth or decline will often be needed. Studies employing such models have primarily appeared in life span developmental and educational psychology.

If the process is believed to be accelerating (or decelerating), then a quadratic term can be added to Equation (25) at Level 1:

$$Y_{ti} = \beta_{0i} + \beta_{1i}(Time)_{ti} + \beta_{2i}(Time)_{ti}^2 + e_{ti} \tag{26}$$

For example, Huttenlocher, Haight, Bryk, Seltzer, and Lyons (1991) used this type of model to capture the growth in children's vocabulary between 12 and 26 months of age in which each child's vocabulary is rapidly accelerating in size. The *Time* predictor variable was centered at 12 months, $Time_C = Time - 12$, a developmental point where outside observers do not detect vocabulary in children. This is the functional 0-point for *Time*. In Equation (26), the β_{0i} term is the predicted vocabulary for child i at 12 months ($Time_C = 0$), the β_{1i} term is the predicted linear rate of growth of child i at 12 months, and β_{2i} is related to the rate of acceleration in vocabulary growth of child i. Such models can be estimated using all MLM and SEM software.

Alternatively, various forms of relationship may be approximated with piecewise models (Raudenbush & Bryk, 2002; Singer & Willett, 2003). For example, Bahrick (1979) found that memory for street names in one's former college town declines rapidly after leaving the location for a period of about 10 years and then levels off. This could be represented by an initial linear piece describing the initial rapid 10-year decline, followed by a second piece beginning at 10 years in which the rate of decline was very low. Piecewise models are also useful for representing known life transitions that occur in the middle of a study, such as when children are measured each year in Grades 1–8, and the end of Grade 5 represents a common known point of transition (e.g., from elementary to middle school; see Wu, West, & Hughes, 2008, for an example involving the transition of retained and promoted children from first to second grade). In other cases (e.g., the transition point between Piagetian stages), the mean

and variance of the transition points can be estimated for the children (Cudeck & Klebe, 2002).

Finally, some processes are believed to lead to initial rapid increases or decreases whose rate of change then declines so that the measure eventually approaches a final level (asymptote). A pattern of rapid decline followed by leveling off at a lower level might be seen in levels of anxiety or physiological indicators (e.g., cortisol), characterizing the recovery from a stressful event over several minutes or even a few days. Such processes can be captured by multilevel exponential models of growth or decline to an asymptote. Cudeck (1996) presented an extensive discussion of such models for estimating multilevel exponential models. These models require special software, such as SAS PROC NLMIXED, that estimates a variety of nonlinear forms of growth.

Multilevel Structural Equation Models

The multilevel models we have considered have all been in either the multiple regression or growth curve modeling traditions. Multilevel models that permit researchers to consider hypothesized measurement (confirmatory factor analysis) models, path (structural) models that hypothesize a network of relationships between multiple variables, or both have also been developed (Goldstein & McDonald, 1988; Lee, 1990; Longford & Muthén, 1992; Muthén & Satorra, 1995). Some of these multilevel extensions of multivariate techniques are beginning to be utilized in psychological research (e.g., Cheung & Au, 2005; Reise, Ventura, Nuechterlein, & Kim, 2005; Zimprich, Perren, & Hornung, 2005).

Multilevel confirmatory factor analysis (CFA) can be used to investigate a hypothesized factor structure at both Levels 1 and 2. Figure 3 shows a hypothetical example based on ideas by Zimprich et al. (2005). A researcher hypothesized a two-level CFA model for a 10-item scale used to measure extroversion. The scale is given to 50 classrooms of 20 students each. Suppose previous research had shown that Items 1, 2, and 3 are heavily influenced by social desirability, whereas Items 4–10 are not influenced by social desirability. As depicted in Figure 4, Items 1–3 at Level 1 are allowed to load on the social desirability construct. All of the Items 1–10 are allowed to load on extroversion. At Level 2, all of the items are allowed to load on extroversion. The two-level CFA model partitions the latent trait into two components: introversion at Level 1 can vary from student

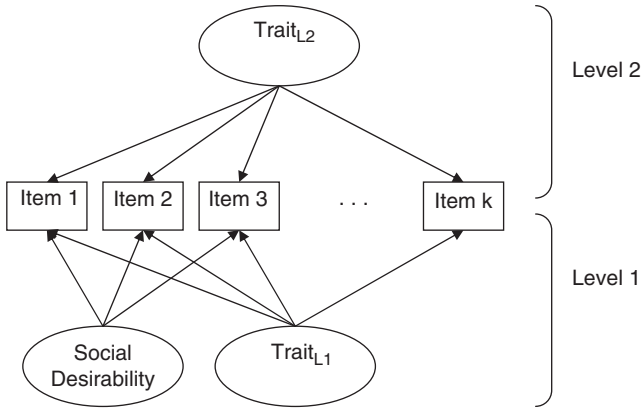

Figure 4
Illustration of multilevel confirmatory factor analysis model. At Level 1, all items are indicators of the individual's level of introversion (Trait$_{L1}$). Items 1–3 are also influenced by social desirability. At Level 2, all items are indicators of the class's level of introversion.

to student (between individuals); introversion at Level 2 can vary from classroom to classroom (e.g., students enrolled in math vs. public speaking classes). The second latent factor only exists at Level 1 and captures the social desirability of each individual on the three sensitive items. In this model, the influences of extroversion at both the individual and at the classroom level are neatly captured, and the potential artifactual influence of individual social desirability is removed from the sensitive items. Thus, both group-level and individual-level influences on the extroversion items can be studied.

Multilevel path models can also be studied. Consider again the stress–alcohol consumption example. Suppose the researchers had assessed the type of primary romantic relationship for each individual at the beginning of the study. Imagine that the researchers hypothesized that individuals in a conflicted (rather than nonconflicted) primary romantic relationship (Level 2) would experience days with very low and very high levels of daily stress (Level 1). Days with higher than average levels of stress would, in turn, lead to greater daily alcohol consumption (Level 1). Thus, we have a simple path model representing a mediational relationship: relationship type (Level 2) → daily stress (Level 1) → daily alcohol consumption (Level 1). Models proposing paths between variables at Level 1, between variables at Level 2, between Level 2 and Level 1 variables,

and between Level 1 and Level 2 variables (e.g., a mentally ill child affecting family functioning) may be proposed. MacKinnon (2008) described methods for testing several mediational effects within both the MLM and SEM frameworks, and Preacher, Zyphur, and Zhang (2010) provided an extensive discussion of the full range of multilevel mediational models within the SEM framework.

Multilevel confirmatory factor analysis and path models can be estimated using several major SEM software packages (e.g., EQS, LISREL, Mplus). One complexity arises because the standard measures of fit used in SEM may be problematic in multilevel models since overall fit is heavily determined by the Level 1 model. Ryu and West (2009) described relatively simple procedures and offer a SAS macro computer program that computes fit indices separately at each level, thereby avoiding this problem.

SUMMARY AND CONCLUSION

Multilevel statistical models offer appropriate procedures for addressing dependency in data and for testing interesting new hypotheses about relationships at the same or different levels in the hierarchy. Basic multilevel procedures are already being commonly utilized by personality researchers in the experience sampling and longitudinal growth areas; other possible areas of application are just beginning to appear in the literature. This article reviewed the specification of models at Levels 1 and 2, noting in particular the care that must be employed in interpreting coefficients. It also considered issues associated with trends or cycles in data collected over time on persons, often ignored in some areas of personality research. The advantages and complexity of different forms of centering in multilevel research were considered, noting that centering within person and restoring the person means (chronic level) at Level 2 of the analysis can neatly partition the data into within-person and between-persons effects. Some consequential differences between MLM and standard OLS regression in model estimation and model comparison were discussed. The often neglected issues of model and data evaluation in MLM were considered. Finally, some of the newer developments that address more complex data structures, funny dependent variables, nonlinear growth, and multilevel structural equation models were briefly presented. Our hope is that this article

will both serve as an introduction to multilevel analysis for new personality researchers and help expand the horizons of more experienced researchers with respect to the areas of application, types of multilevel models, and issues that arise in multilevel analysis.

REFERENCES

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

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno (1990) survey of Ph.D. programs in North America. *American Psychologist*, **63**, 32–50.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Akademiai Kiado.

Armeli, S., Carney, M. A., Tennen, H., Affleck, G., & O'Neil, T. P. (2000). Stress and alcohol use: A daily process examination of the stressor-vulnerability model. *Journal of Personality and Social Psychology*, **78**, 979–994.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, **108**, 296–308.

Bauer, D. J., & Curran, P. J. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, **40**, 373–400.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.

Beretvas, S. N. (2008). Cross-classified random effects models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 161–197). Charlotte, NC: Information Age.

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

Biesanz, J. C., West, S. G., & Kwok, O. M. (2003). Personality over time: Methodological approaches to the study of short-term and long-term development and change. *Journal of Personality*, **71**, 905–942.

Biesanz, J. C., West, S. G., & Millevoi, A. (2007). What do you learn about someone over time? The relationship between length of acquaintance and consensus and self-other agreement in judgments of personality. *Journal of Personality and Social Psychology*, **92**, 119–135.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley.

Bosker, R. J., Snijders, T. A. B., & Guldemond, H. (2003). *PINT (Power in two-level designs) user's manual*. Oxford, UK: Author. Available from http://stat.gamma.rug.nl/

Browne, W. J., Golalizadeh-Lahi, M., & Parker, R. M. A. (2009). *A guide to sample size calculations for random effect models via simulation and the*

MLPowSim software package. University of Bristol, UK. Retrieved October 29, 2010, from http://www.cmm.bristol.ac.uk/learning-training/multilevel-models/samples.shtml

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

Cheung, M. W. L., & Au, K. (2005). Applications of multilevel structural equation modeling to cross-cultural research. *Structural Equation Modeling*, **12**, 598–619.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Conner, T. S., Feldman-Barrett, L., Tugade, M. M., & Tennen, H. (2007). Idiographic personality: The theory and practice of experience sampling. In R. W. Robins, R. C. Fraley, & R. R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 79–96). New York: Guilford Press.

Coxe, S. J., West, S. G., & Aiken, L. S. (in press). Generalized linear models. In T. Little (Ed.), *Oxford handbook of quantitative methods*. New York: Oxford University Press.

Cramer, P., & Jones, C. J. (2007). Defense mechanisms predict differential lifespan change in self-control and self-acceptance. *Journal of Research in Personality*, **41**, 841–855.

Cudeck, R. (1996). Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate Behavioral Research*, **31**, 371–403.

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed effects models for repeated measures data. *Psychological Methods*, **7**, 41–63.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, **38**, 529–568.

de Leeuw, J., & Meijer, E. (Eds.). (2008). *Handbook of multilevel analysis*. New York: Springer.

Enders, C. K. (2005). Maximum likelihood estimation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1164–1170). Chichester, UK: Wiley.

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.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, **73**, 43–56.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Arnold.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, **53**, 455–467.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics*, **9**, 226–252.

Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research*, **42**, 609–629.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Hox, J. J., & Maas, C. J. (2001). The accuracy of multilevel structural equation modeling with pseudobalanced groups and small samples. *Structural Equation*, **8**, 157–174.

Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology*, **27**, 236–248.

Jones, C. J., & Meredith, W. (1996). Patterns of personality change across the life span. *Psychology and Aging*, **11**, 57–65.

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

Krahé, B. (1992). *Personality and social psychology: Towards a synthesis*. Thousand Oaks, CA: Sage.

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

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, **30**, 1–21.

Kwok, O. M., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, **42**, 557–592.

Langford, I. H., & Lewis, T. (1998). Outliers in multilevel data. *Journal of the Royal Statistical Society, Series A*, **161**, 121–160.

Lee, S. Y. (1990). Multilevel analysis of structural equation models. *Biometrika*, **77**, 763–772.

Lesaffre, E., & Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

Lewis, T., & Langford, I. H. (2001). Outliers, robustness and the detection of discrepant data. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modeling of health statistics* (pp. 75–91). New York: Wiley.

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, **57**, 581–597.

Lucas, R. E., Clark, A. E., Georgellis, Y., & Diener, E. (2003). Reexamining adaptation and the set point model of happiness: Reactions to changes in marital status. *Journal of Personality and Social Psychology*, **84**, 527–539.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, **13**, 203–229.

Luo, W., & Kwok, O. M. (2009). Impacts of ignoring a crossed factor in analyzing multilevel data with cross-classified structures: A Monte Carlo study. *Multivariate Behavioral Research*, **44**, 182–212.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Erlbaum.

McArdle, J. J. (2009). Latent variable modeling of differences and changes in longitudinal data. *Annual Review of Psychology*, **60**, 577–605.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, **38**, 115–142.

McLean, K. C., & Fournier, M. A. (2008). The content and processes of autobiographical reasoning in narrative identity. *Journal of Research in Personality*, **42**, 527–545.

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

Meyers, J., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, **41**, 473–497.

Mok, M. (1995). Sample size requirements for 2-level designs in educational research. *Multilevel Modeling Newsletter*, **7**, 11–15.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, **9**, 599–620.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, **25**, 267–316.

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

Nezlek, J. B. (2007). Multilevel modeling in personality research. In R. W. Robins, R. C. Fraley, & R. R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 502–522). New York: Guilford Press.

Oishi, S., Diener, E., Choi, D., Kim-Prieto, C., & Choi, I. (2007). The dynamics of daily events and well-being across cultures: When less is more. *Journal of Personality and Social Psychology*., **93**, 685–698.

Oishi, S., Lun, J., & Sherman, G. D. (2007). Residential mobility, self-concept, and positive affect in social interactions. *Journal of Personality and Social Psychology*, **93**, 131–141.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, **31**, 437–448.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, **15**, 209–233.

Rasbash, J., & Browne, W. J. (2008). Non-hierarchical multilevel models. In J. de Leeuw & E. Meyer (Eds.), *Handbook of multilevel analysis* (pp. 301–334). New York: Springer.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, **19**, 337–350.

Raudenbush, S. W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational and Behavioral Statistics*, **18**, 321–349.

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y., Congdon, R., & du Toit, M. (2004). *HLM6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: SSI.

Reis, H. T., & Gable, S. L. (2000). Event-sampling and other methods for studying daily experience. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 190–222). New York: Cambridge University Press.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, **84**, 126–136.

Roberts, J. K. (2007, April). *Group dependency in the presence of small intraclass correlation coefficients: An argument in favor of not interpreting the ICC*. Paper presented at the American Educational Research Association Meeting, Chicago. Retrieved October 29, 2010, from http://www.hlm-online.com/papers/depend.pdf

Ryu, E. (2004). *Effects of small group sizes on the estimation of multilevel models: A Monte Carlo study*. Unpublished master's thesis, Arizona State University.

Ryu, E., & West, S. G. (2009). Level-specific evaluation of model fit in multilevel structural equation modeling. *Structural Equation Modeling*, **16**, 583–601.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Shi, L., & Chen, G. (2009). Influence measures for general linear models with correlated errors. *American Statistician*, **63**, 40–42.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, **86**, 420–428.

Shrout, P. E., Herman, C. E., & Bolger, N. (2006). The costs and benefits of practical and emotional support on adjustment: A daily diary study of couples experiencing acute stress. *Personal Relationships*, **13**, 115–134.

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

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Boca Raton, FL: Chapman & Hall/CRC.

Snijders, T. A. B., & Berkhof, J. (2008). Diagnostic checks for multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 140–175). New York: Springer.

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, **22**, 342–363.

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

Snyder, M., & Ickes, W. (1985). Personality and social behavior. In G. Lindsay & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 2, pp. 883–947). New York: Random House.

Stoel, R. D., Garre, F. G., Dolan, C., & van den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods*, **11**, 439–455.

Tamir, M., John, O. P., Srivastava, S., & Gross, J. J. (2007). Implicit theories of emotion: Affective and social outcomes across a major life transition. *Journal of Personality and Social Psychology*, **92**, 731–744.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental personality designs: Analyzing categorical by continuous variable interactions. *Journal of Personality*, **64**, 1–48.

West, S. G., Aiken, L. S., Wu, W., & Taylor, A. B. (2007). Multiple regression: Applications of the basics and beyond in personality research. In R. W. Robins, R. C. Fraley, & R. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 573–601). New York: Guilford Press.

West, S. G., & Hepworth, J. T. (1991). Statistical issues in the study of temporal data: Daily experiences. *Journal of Personality*, **59**, 609–662.

Wu, W., West, S. G., & Hughes, J. N. (2008). Effect of retention in first grade on children's achievement trajectories over four years: A piecewise growth analysis using propensity score matching. *Journal of Educational Psychology*, **100**, 727–740.

Zimprich, D., Perren, S., & Hornung, R. (2005). A two-level confirmatory factor analysis of a modified Rosenberg self-esteem scale. *Educational and Psychological Measurement*, **65**, 465–481.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Supplement A. Data Generation

Supplement B(1). SAS Syntax to Estimate Models A–E

Supplement B(2). SPSS Syntax to Estimate Models A–E

Supplement C. SAS PROC MIXED syntax to create a residual file that contains the random intercept and slope estimates.

Supplement D(1). SAS Syntax to Create Q-Q Plot Using the Expected and Observed Mahalanobis Distances From the HLM Level 2 Residual File

Supplement D(2). SPSS Syntax to Create Q-Q Plot Using the Expected and Observed Mahalanobis Distances From the HLM Level 2 Residual File

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.