CHAPTER 30

Multilevel Analysis in the Study of Crime and Justice

BRIAN D. JOHNSON

"The most pervasive fallacy of philosophic thinking goes back to neglect of context"
(John Dewey, 1931)

Neither criminal behavior nor society's reaction to it occurs in a social vacuum – for this reason, criminology, as a discipline, is inherently a multilevel enterprise. Individual criminal behavior is influenced by larger social, political, and environmental factors, as are the decisions of various actors in the criminal justice system. Classroom and school characteristics affect adolescent development, misconduct and delinquency (Beaver et al. 2008; Osgood and Anderson 2004; Stewart 2003). Family and neighborhood characteristics influence the likelihood of victimization and offending, as well as postrelease recidivism and fear of crime (Nieuwbeerta et al. 2008; Wilcox et al. 2007; Lauritsen and Schaum 2004; Kubrin and Stewart 2006; Wyant 2008; Lee and Ulmer 2000). Police department, precinct, and neighborhood factors affect police arrest practices, use of force, and clearance rates (Smith 1986; Sun et al. 2008; Lawton 2007; Pare et al. 2007; Eitle et al. 2005; Terrill and Reisig 2003). Judge characteristics and court contexts affect individual punishment decisions (Britt 2000; Ulmer and Johnson 2004; Johnson 2006; Wooldredge 2007), and prison environments are tied to inmate misconduct, substance use, and violence (Camp et al. 2003; Gillespie 2005; Huebner 2003; Wooldredge et al. 2001). Although these examples cover a diverse array of criminological topics, they all share a common analytical quality – each involves data that are measured across multiple units of analysis. When this is the case, multilevel models offer a useful statistical approach for studying diverse issues in crime and justice. As Table 30.1 demonstrates, recent years have witnessed an abundance of multilevel studies across a variety of topics in criminology and criminal justice.

This chapter provides a basic introduction to the use of multilevel statistical models in the field. It begins with a conceptual overview explaining what multilevel models are and why they are necessary. It then provides a statistical overview of basic multilevel models, illustrating their application using punishment data from federal district courts. The chapter concludes with a discussion of advanced applications and common concerns that arise in the context of multilevel research endeavors.

A.R. Piquero and D. Weisburd (eds.), *Handbook of Quantitative Criminology*,
DOI 10.1007/978-0-387-77650-7_30,

TABLE 30.1. Recent examples of multilevel studies published in *Criminology* from 2005 to 2008

References	Topic
Xie and McDowall (2008)	Victimization and residential mobility
Schreck et al. (2008)	Violent offender and victim overlap
Johnson et al. (2008)	Federal guidelines departures
Xie and McDowall (2008)	Residential turnover and victimization
Mears et al. (2008)	Social context and recidivism
Zhang et al. (2007)	Crime reporting in China
Kreager (2007)	School violence and peer acceptance
Wilcox et al. (2007)	Guardianship and burglary victimization
Chiricos et al. (2007)	Labeling and felony recidivism
Osgood and Schreck (2007)	Stability and specialization in violence
Rosenfeld et al. (2007)	Order-maintenance policing and crime
Bernburg and Thorlindsson (2007)	Community structure and delinquency
Warner (2007)	Social context and calls to police
Hay and Forrest (2006)	The stability of self-control
Doherty (2006)	Self-control, social bonds, and desistence
Griffin and Wooldredge (2006)	Sex disparities in imprisonment
Sampson et al. (2006)	Marriage and crime reduction
Ulmer and Bradley (2006)	Trial penalties
Johnson (2006)	Judge and court context in sentencing
Kubrin and Stewart (2006)	Neighborhood context and recidivism
Simons et al. (2005)	Collective efficacy, parenting and delinquency
Slocum et al. (2005)	Strain, offending and drug use
Wright and Beaver (2005)	Parental influence and self-control
Bontrager et al. (2005)	Race and adjudicated guilt
Kleck et al. (2005)	Perceptions of punishment
Johnson (2005)	Sentencing guidelines departures

CONCEPTUAL OVERVIEW

Multilevel statistical models are necessitated by the fact that social relationships exist at several different levels of analysis that jointly influence outcomes of interest. Smaller units of analysis are often "nested" within one or more larger units of analysis. For instance, students are nested within classrooms and schools, offenders as well as victims are nested within family and neighborhood environments, and criminal justice personnel are nested within larger community and organizational contexts. In each of these cases, the characteristics of some larger context are expected to influence individual behavior. This logic can be extended to any situation involving multiple levels of analysis including, but not limited to, individuals, groups, social networks, neighborhoods, communities, counties, states, and even countries. Moreover, longitudinal research questions often involve multilevel data structures, with repeated measures nested within individuals or with observations nested over time (e.g., Horney et al. 1995; Slocum et al. 2005; Rosenfeld et al. 2007). Other common applications of multilevel analysis include twin studies with paired or clustered sibling dyads (e.g. Wright et al. 2008; Taylor et al. 2002) and meta-analyses that involve multiple effect sizes nested within the same study or dataset (e.g. Raudenbush 1984; Goldstein et al. 2000). Regardless of the level of aggregation or "nesting," though, the important point is that criminological enterprises often involve data that span multiple levels of analysis. In fact, given the complexity of our social world, it can be difficult to identify topics of criminological interest that are not characterized by multiple spheres of social influence. When these multiple influences are present, multilevel statistical models represent a useful and even necessary tool for analyzing a broad variety of criminological research questions.

A Model by any Other Name?

Given the complexity surrounding multilevel models, it is useful to distinguish up front what can sometimes be a confusing and inexact argot. Various monikers are used to describe multilevel statistical models (e.g. multilevel models, hierarchical models, nested models, mixed models). Although this nomenclature is often applied interchangeably, there can be subtle but important differences in these designations. Multilevel modeling is used here as a broad, all-encompassing rubric for statistical models that are explicitly designed to analyze and infer relationships for more than one level of analysis.[1] The language of multilevel models is further complicated by the fact that there are various different software packages that can be used to estimate multilevel models, some of which are general statistical packages (e.g. SAS, STATA) while others are specialized multilevel programs (e.g. HLM, MLwin, aML).[2]

Additional confusion may derive from the fact that scholars often use terminology, such as ecological, aggregate, and contextual effects, interchangeably despite important differences in their meaning. *Ecological effects*, or group-level effects, can refer to any group-level influence that is associated with the higher level of analysis. Group-level effects can take several forms. First, *aggregate effects* (sometime referred to as *analytical* or *derived* variables) are created by aggregating individual level characteristics up to the group-level of analysis (e.g., percent male in a school). These are sometimes distinguished from *structural effects* that are also derived from individual data but capture relational measures among members within a group (e.g., density of friendship networks) (see Luke 2004: 6). To complicate matters, when individual-level data are aggregated to the group-level, they can exert two distinct types of influence – first, they can exert *compositional effects*, which reflect group-differences that are attributable to variability in the constitution of the groups – between-group differences may simply reflect the fact that groups are made up of different types of individuals. Second, they can exert *contextual effects*, which represent influences above and beyond differences that exist in group composition. Contextual effects are sometimes referred to as *emergent properties* because the collective exerts a synergistic influence that is unique to the group aggregation

[1] Hierarchical or nested models, for instance, technically refer to data structures involving exact nesting of smaller levels of analysis within larger units. Multilevel data, however, can also be nonnested, or "cross-classified", in ways that do not follow a neat hierarchical ordering. Data might be nested within years and within states at the same time, for example, with no clear hierarchy to "year" and "state" as levels of analysis (Gelman and Hill 2007: 2). Similarly, adolescents might be nested within schools and neighborhoods with students from the same neighborhood attending different schools, or convicted terrorists might be nested both within terrorist groups and the district courts in which they are punished. Although these cases clearly involve multilevel data, they are not hierarchical in a technical sense. Similarly, "mixed models" technically refer to statistical models containing both "fixed" and "random" effects. Although this is often the case in multilevel models, it is not necessarily so; the broader rubric multilevel modeling is preferred, therefore, to capture the variety of models designed to incorporate data across multiple units of analysis.

[2] A useful and detailed review of the strengths and limitations of numerous software programs that provide for the estimation of multilevel models is provided by the Centre for Multlevel Modeling at the University of Bristol at http://www.cmm.bris.ac.uk/learning-training/multilevel-m-software/index.shtml.

and is not present in the individual constituent parts.[3] Although the term "contextual effect" is sometimes used as a broader rubric for any group-level influence, the narrower definition provided here is often useful for distinguishing among types of ecological influences that can be examined in multilevel models. Finally, *global effects* refer to structural characteristics of the collective itself that are not derived from individual data, but rather reflect measures that are specific to the group (e.g., physical dilapidation of the school). These and other commonly used terms in multilevel analysis are summarized in Table 30.2.[4]

Theoretical Rationales for Multilevel Models

The need for multilevel statistical models is firmly rooted in both theoretical and methodological rationales. Multilevel models are extensions of traditional regression models that account for the structuring of data across aggregate groupings, that is, they explicitly account for the nested nature of data across multiple levels of analysis. Because our social world is inherently multilevel, theoretical perspectives that incorporate multiple levels of influence in the study of crime and justice are bound to improve our ability to explain both individual criminal behavior and society's reaction to it.

Figure 30.1 presents a schematic of a hypothetical study examining the influence of low self-control on delinquency in a sample of high school students. Imagine that self-control is measured at multiple time points for the same sample of students. In such a case, multiple measures of self-control would be nested within individual students, and individual students would be nested within classrooms, which are nested within schools. The lowest level of analysis would be within-individual observations of self-control, and the highest level of analysis would be school-level characteristics. Ignoring this hierarchical data structuring, then, is likely to introduce omitted variable bias of a large-scale theoretical nature.

Moreover, many theoretical perspectives explicitly argue that micro-level influences will vary across macro-social contexts. For instance, racial group threat theories (Blumer 1958; Liska 1992) predict that the exercise of formal social control will vary in concert with large or growing minority populations. Testing theoretical models that explicitly incorporate variation in micro-effects across macro-theoretical contexts, therefore, offers an important opportunity to advance criminological knowledge. Assuming that theoretical influences operate at a single level of analysis is likely to provide a simplistic and incomplete portrayal of the complex criminological social world.

Moreover, inferential problems can emerge when data are used to draw statistical conclusions across levels of analysis. For instance, in his classic study of immigrant literacy,

[3] Philosophical discourse on emergent properties dates all the way back to Aristotle, but was perhaps most lucidly applied by John Stuart Mill. He argued that the human body in its entirety produces something uniquely greater than its singular organic parts, stating that "To whatever degree we might imagine our knowledge of the properties of the several ingredients of a living body to be extended and perfected, it is certain that no mere summing up of the separate actions of those elements will ever amount to the action of the living body itself" (Mill 1843). Contextual effects models have long been applied in sociology and related fields (e.g., Firebaugh 1978; Blalock 1984), but these applications differ from multilevel models in that the latter are more general formulations that specifically account for residual correlation within groups and explicitly provide for examination of the causes of between-group variation in outcomes.

[4] Table 30.2 is partially adapted from Diez Roux (2002), which contains a more detailed and elaborate glossary of many of these terms.

TABLE 30.2. Glossary of multilevel modeling terminology

Terminology	Definition
Aggregate variable	Ecological variable created by aggregating the individual properties of lower level measures up to the group level of analysis. Sometimes also referred to as "Derived" or "Analytical" variables.
Atomistic fallacy	The fallacy, also referred to as the Individualistic Fallacy, that results when faulty inferences for macro-level group relationships are drawn using micro-level individual data. See Ecological fallacy
Compositional effects	Between group differences in outcomes that are attributable to differences in group composition, or in the different individuals of which the groups are composed
Contextual analysis	Early analytical approach designed to investigate the effects of aggregate characteristics of the collective by including aggregate variables along with individual variables in traditional regression models.
Contextual effects	Macro-level influences exerted by aggregate variables above and beyond those attributable to compositional differences in groups, but sometimes the term is used to refer to any group level effects.
Contextual effects model	Statistical model that include individual characteristics and the aggregates of the individual characteristics in the same model in order to assess the influence of contextual effects on individual outcomes.
Cross-level interaction	A statistical interaction between higher and lower order variables, usually attempting to explain variation in the effects of lower level measures across higher level groupings.
Cross-classified model	A multilevel statistical model for analyzing data that is cross-nested in two or more higher levels of analysis which are not strictly hierarchical in structure. Also referred to as cross-nested models.
Ecological fallacy	The fallacy that results when faulty inferences for individual level relationships are made using group level data. See Atomistic fallacy.
Ecological variable	A broad term for any higher order group level variable, including aggregate, structural, and global measures. Sometimes referred to as a Group Level, Macro Level, or Level 2 variable.
Empirical bayes estimates	Estimates for group level parameters that are optimally weighted to combine information from the individual group itself with information from other similar groups in the data.
Fixed effects	Regression coefficients (or intercepts) that are not allowed to vary randomly across higher level units. These are sometimes referred to as fixed coefficients. See Random effects.
Fixed effects models	Statistical models in which all effects or coefficients are fixed. Often this refers to the case where a dummy variable is included for each higher level unit to remove between-group variation in the outcome.
Global variable	A group level variable that unlike aggregate variables has no individual analogue. Global (or integral) variables refer to characteristics that are uniquely defined at the higher level of analysis.
Group level variable	An alternative name for ecological variables that measure any group level characteristic. Sometimes referred to as Level 2 variables. See Individual level variable.
Hierarchical (linear) model	A multilevel model for analyzing data that is nested among two or more hierarchies. Hiearchical models technically assume that data are strictly nested across levels of analysis, although this term may refer to multilevel models generally.
Individual level variable	A variable that characterizes individual attributes or refers to individual level constructs. Sometimes referred to as Level 1 variables. See Group level variable.

(continued)

TABLE 30.2. (continued)

Terminology	Definition
Intraclass correlation	The proportion of the total variance in the outcome that exists between groups or higher level units rather than within groups or higher level units.
Mixed model	A multilevel model containing both fixed and random coefficients. Some regression coefficients are allowed to vary randomly across higher level units while other regression coefficients are specified as fixed coefficients.
Multilevel analysis	An analytical approach for simultaneously analyzing both individual and group level effects when data is measured at two or more levels of analysis with lower level (micro) observations nested within higher level (macro) units.
Multilevel model	A statistical model used in multilevel analysis for analyzing data that is measured at two or more levels of analysis, including but not limited to hierarchical linear models, hiearchical nonlinear models, and cross-classified models.
Population average estimates	Estimates for nonlinear multilevel models that provide the marginal expectation of the outcome averaged across all random effects rather than after controlling for random effects. See Unit-specific estimates
Random coefficient model	A multilvel statistical model in which the individual level intercept and regression coefficients are allowed to have randomly varying effects across higher level units of analysis. See Random intercept model.
Random effects	Regression coefficients (or intercepts) that are allowed to vary randomly across higher level units. These are sometimes referred to as random intercepts or random coefficients. See Fixed effects.
Random intercept model	A multilevel statistical model in which the individual level intercept is allowed to vary randomly across higher level units of analysis, but the individual level coefficients are assumed to have constant effects. See Random coefficient model.
Unit-specific estimates	Estimates for nonlinear multilevel models that are conditional on higher level random effects. Unit-specific models provide individual estimates controlling for rather than averaging across random effects. See Population average estimates.
Variance components	Model parameters (sometimes referred to as random effects) that explicitly capture both within-group and between-group variability in outcomes. Each level of analysis in a multilevel model has its own variance component.

Robinson (1950) examined the correlation between aggregate literacy rates and the proportion of the population that was immigrant at the state level. He found a substantial positive correlation between percent immigrant and the literacy rate ($r = 0.53$). Yet, when individual level-data on immigration and literacy were separately examined, the correlation reversed and became negative ($r = -0.11$). Although individual immigrants had lower literacy, they tended to settle in states with high native literacy rates, thus confounding the individual and aggregate relationships. This offers an example of the *ecological fallacy*, or erroneous conclusions involving individual relationships that are inferred from aggregate data. As Peter Blau (1960: 179) suggested, aggregate studies are limited because they cannot "separate the consequences of social conditions from those of the individual's own characteristics for his behavior, because ecological data do not furnish information about individuals except in the aggregate."

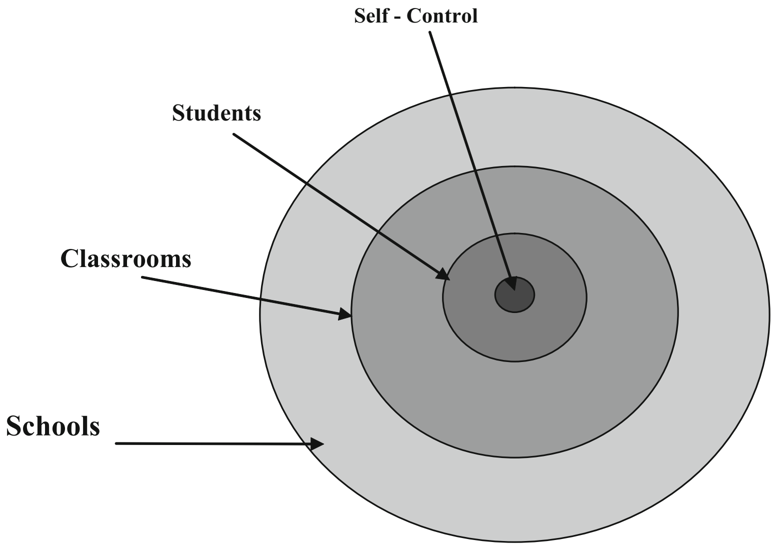


FIGURE 30.1. The hierarchical nature of multiple units of analysis in multilevel models.

The same mistake in statistical inference can occur in the opposite direction. The *atomistic (or individualistic) fallacy* occurs when aggregate relationships are mistakenly inferred from individual-level data. Because associations between two variables at the individual level may differ from associations for analogous variables at a higher level of aggregation, aggregate relationships cannot be reliably inferred from individual-level data. For instance, social disorganization theory would predict that crime rates across neighborhoods are related to mobility rates because high population turnover reduces informal social control at the neighborhood level. Because this prediction refers to neighborhoods as the unit of analysis, though, one would risk serious inferential error testing this group-level hypothesis with individual-level data. For instance, one could not test the theory by examining whether or not individuals who move residences have higher criminal involvement. To do so would be to commit the atomistic fallacy. This reflects the fact that variables aggregated up from individual-level data often have unique and independent *contextual* effects. Moving to a new residence represents a different causal pathway than living in a neighborhood with high rates of residential mobility.[5]

Because of the inherent difficulties in making statistical inferences across different levels of analysis, a preferred approach is to use multilevel analytic procedures to simultaneously incorporate individual- and group-level causal processes. Multilevel models explicitly provide for this type of statistical analysis. The difficulty is in distinguishing among the different types of individual and ecological influences that are of theoretical interest and then specifying the statistical model to properly estimate these effects.

[5] Two related, but distinct problems of causal inference are the *psychologistic fallacy*, which can occur when individual level data are used to draw inferences without accounting for confounding ecological influences, and the *sociologistic fallacy*, which may arise from the failure to consider individual level characteristics when drawing inferences about the causes of group variability. The pyschologistic fallacy results from a failure to adequately consider *contextual effects*, whereas the sociologistic fallacy results from a failure to capture *compositional effects*.

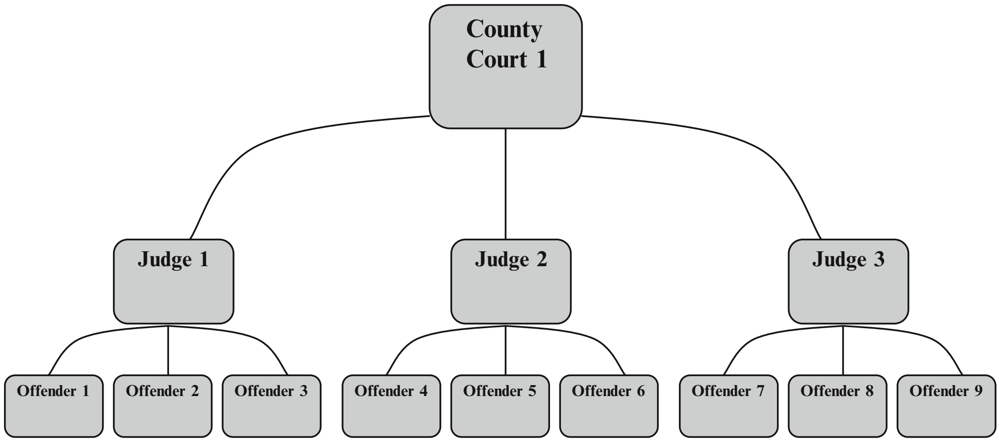


FIGURE 30.2. The nesting of multilevel data across levels of analysis.

Multilevel Modeling and Hypothesis Testing

There are also persuasive statistical reasons for engaging in multilevel modeling such as providing improved parameter estimates, corrected standard errors, and conducting more accurate statistical significance tests. Utilizing traditional regression models for multilevel data presents several problems. Figure 30.2 presents a second example of hierarchical data where individual criminal offenders are nested within judges and county courts. Because several offenders are sentenced by the same judge and several judges share the same courtroom environment, statistical dependencies are likely to arise among clustered observations. When individual data is nested within aggregate groups, observations within clusters are likely to share unaccounted-for similarities. If, for instance, some judges are "hanging judges" while others are "bleeding-heart liberals," then offenders sentenced by the former will have sentences that are systematically harsher than offenders sentenced by the latter. Statistically speaking, the residual errors will be correlated, systematically falling above the regression line for the first judge and below it for the second. Because one of the assumptions of ordinary regression models is that residual errors are independent, such systematic clustering would violate this core model assumption. The consequence of this violation is that standard errors will be *underestimated* by the ordinary regression model. Statistical significance tests will therefore be too liberal, risking Type I inferential errors in which the null hypothesis is falsely rejected even when true in the population. Multilevel statistical models are needed to account for statistical dependencies that occur among clusters of hierarchically organized data.

A related problem is that statistical signifiance tests in ordinary regression models utilize the wrong degrees of freedom for ecological predictors in the model. Traditional regression models fail to account for the fact that hierarchically structured data are characterized by different sample sizes at each level of analysis. For example, with data on 1,000 students nested within 50 schools, there would be an individual-level sample size of 1,000 observations, but a school-level sample size of only 50 observations. This means that statistical significance tests for school-level predictors need to be based on degrees of freedom that reflect the number of schools in the data, not the number of students. Statistical significance tests in ordinary regression models fail to recognize this important distinction. The consequence is that the amount of statistical power available for testing school-level predictors will be exaggerated. The number of degrees of freedom for statistical significance tests needs to be adjusted for the number of aggregate units in the data – multilevel models provide these adjustments.

A third advantage of multilevel models over ordinary regression models is that they allow for the modeling of heterogeneity in regression effects. The single-level regression model assumes *de facto* that individual predictors exert the same effect in each aggregate grouping. Multilevel models, on the other hand, explicitly allow for variation in the effects of individual predictors across higher levels of analysis. Ulmer and Bradley, (2006), for instance, have argued that the effect of trial conviction on criminal sentence varies across courts. This proposition is illustrated in Fig. 30.3, using federal punishment data for a random sample of eight district courts. The ordinary regression model would constrain the effect of trial conviction to be uniform across courts, but Fig. 30.3 clearly suggests intercourt variation in this effect. Multilevel analysis allows for this type of variation to be explicitly incorporated into the statistical model, providing the researcher with a useful tool for better capturing the real-world complexity that is likely to characterize individual influences across criminological contexts.

Other advantages that also characterize multilevel models are that they provide for convenient and accurate tests of cross-level interactions, or moderating effects that involve both individual and ecological variables. For example, the influence of individual socioeconomic status on delinquency might depend on the socioeconomic composition of the school. This conditional relationship could be directly investigated by specifying a cross-level interaction between an individual's SES and the mean SES at the school level.

One final statistical advantage of multilevel models is that they are able to simultaneously incorporate information both within and between groups in order to provide optimally weighted group-level estimates. This is accomplished by combining information from the group itself with information from other similar groups in the data, and it is particularly useful when some groups have relatively few observations. Because groups with smaller sample sizes will have less reliable group means, some regression to the overall grand mean is expected. Utilizing a Bayesian estimation approach, the multilevel model shifts the within-group mean toward the mean for other groups. The more reliable the group mean, the more heavily it is weighted; the less reliable (and the less variability across groups), the more the estimate is shifted toward the overall grand mean for all groups in the data. Thus, estimates for specific groups are based not only on their own within-group data, but also on data from other groups. This process is sometimes referred to as "borrowing power" because within-group estimates benefit from information on other groups, and the estimates themselves are sometimes called "shrinkage estimates" because they "shrink" individual group means toward the grand mean for all groups. The end result is that group-level estimates are optimally weighted to reflect information both within and between groups in the data.[6]

[6] The following equation provides the formula for this weighting process (Raudenbush and Bryk 2002: 46):

$$\hat{\beta}_j^* = \lambda_j \bar{Y}_{\bullet j} + (1 - \lambda_j)\hat{\gamma}_{00}$$

where $\hat{\beta}_j^*$ is the group estimate, which is a product of the individual group mean $\overline{Y}_{\bullet j}$ weighted by its reliability λ_j, plus the overall grand mean $\hat{\gamma}_{00}$ weighted by the complement of the reliability $(1 - \lambda_j)$. If the reliability of the group mean is one, the weighted estimate reduces to the group mean; if it is zero, it reduces to the grand mean. The more reliable the group mean, the more it counts in the multilevel estimate. When the assumptions of the multilevel model are met, this provides the most precise and most efficient estimator of the group mean.

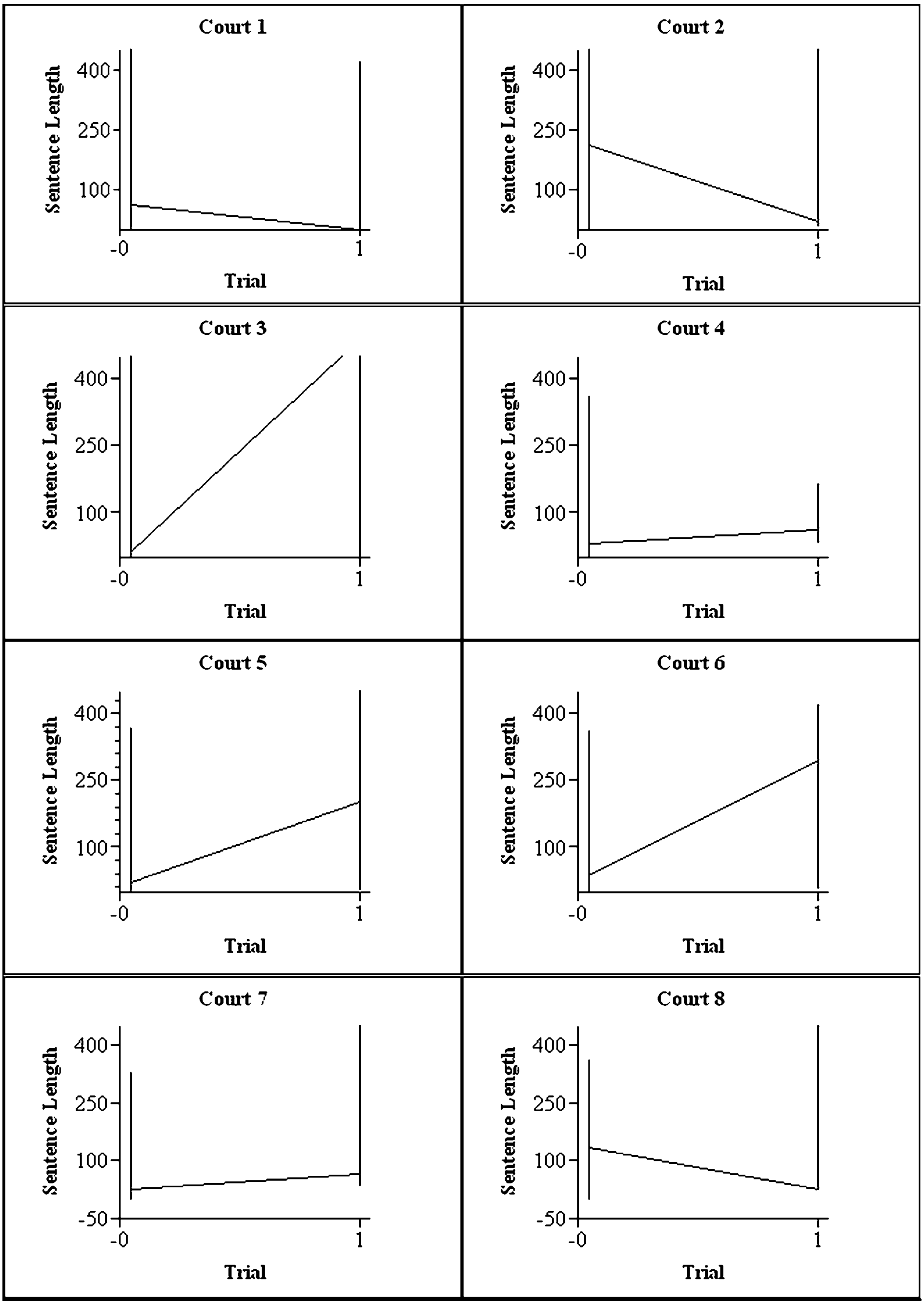


FIGURE 30.3. Variation in trial penalties across federal courts.

Multilevel models, then, provide numerous analytical and statistical advantages over ordinary regression approaches when data are nested across levels of analysis. By providing for the simultaneous inclusion of individual- and group-level information, they better specify the complex relationships that often characterize our social world, and they help overcome common problems of statistical inference associated with reliance on single-level data. Moreover, multilevel models correct for the problematic clustering of observations that may occur with nested data, they provide a convenient approach for modeling both within and between group variability in regression effects, and they offer improved parameter estimates that simultaneously incorporate within and between group information. The remaining discussion provides a basic statistical introduction to the multilevel model along with examples illustrating its application to the study of criminal punishment in federal court.

STATISTICAL OVERVIEW

Multilevel models are simple extensions of ordinary regression models, which account for the nesting of data within higher-order units. It is therefore useful to begin with an overview of the basic regression model in order to demonstrate how the multilevel adaptation builds upon and extends it to the case of multilevel data. For illustrative purposes, examples are provided using United States Sentencing Commission (USSC) data on a random sample of 25,000 convicted federal offenders nested within 89 federal district courts across the US.[7]

From Ordinary Regression to Multilevel Analysis

When faced with multilevel data (e.g. lower-level data that is nested within some higher-level grouping), ordinary regression approaches can take three basic forms. First, individual data can be pooled across groups and analyzed without regard for group structure. This approach ignores important group-level variability and often violates key assumptions of OLS regression such as independent errors. Second, separate unpooled analyses can be conducted within each group. This approach can be useful for examining between-group variability, but it requires relatively large samples for each group and it is cumbersome when the number of groups becomes large. Third, aggregate analysis can also be conducted at the group level alone, but this approach ignores within-group variability and requires a relatively large number of groups for analysis. In each of these cases, traditional regression approaches are unable to incorporate the full range of information available at both the individual and group level of analysis and they may violate important assumptions of the single-level ordinary regression model.

For illustrative purposes, the ordinary regression model is presented in (30.1):

$$Y_i = \beta_0 + \beta_1 X_i + r_i \tag{30.1}$$

where Y_i is a continuous dependent variable, β_0 is the model intercept, β_1 is the effect of the independent variable X_i for individual i, and r_i is the individual-level residual error term. Two key assumptions of the linear regression model are that the relationship between X_i and Y_i can be summarized with a single linear regression line and that all of the residual error terms for individuals in the data are statistically independent of one another. Both of these

[7] These data are drawn from fiscal years 1997 to 2000 and are restricted to the 89 federal districts and 11 circuit courts within the US, with the District of Columbia excluded because it has its own district and circuit court. For more information on the USSC data see Johnson et al. (2008).

assumptions are likely to be violated with multilevel data, the first because the effect of X_i on Y_i might vary by group and the second because individuals within the same group are likely to share unaccounted-for similarities.

Failure to account for the nesting of observations can result in "false power" at both levels of analysis. False power occurs because there is typically less independent information available when observations are clustered together. Consider the difference between (a) data from 50 schools with 20 students each, versus (b) data from 1,000 schools with one student each. The number of students is the same, but if students share similarities within schools, each student provides less unique information in the first sample than in the second. Moreover, there is more unique school-level data in the second sample than in the first. Because ordinary regression models ignore the clustering of individuals within schools, they treat both samples as equivalent. The consequence of this is that the amount of statistical power for the first sample is artificially inflated at both the individual and school level of analysis. Moreover, standard errors for the first sample will be underestimated and significance tests will be too liberal if there are unaccounted-for similarities among students within schools.

The multilevel solution is to add an additional error parameter to the ordinary regression model in order to capture group-level dependencies in the data. The multilevel model is represented by a series of "submodels" that model between-group variation in individual-level parameters as a function of group level processes. A basic two-level random intercept model is presented in (30.2):

$$\begin{aligned} &\text{Level 1} \quad Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \\ &\text{Level 2} \quad \beta_{0j} = \gamma_{00} + u_{0j} \end{aligned} \tag{30.2}$$

where the level one intercept β_{0j} is modeled as an outcome in the level 2 portion of the model. The γ_{00} parameter represents the Level 2 intercept (gammas are substituted for betas at Level 2 for notational convenience) and the u_{0j} parameter represents the new group-level error term, which accounts for group-level dependence. The two-level model specification is presented for simple notational convenience and can be combined into an equivalent single-level model by substituting the Level 2 model in for β_{0j} at Level 1. Doing so produces the combined model in (30.3):

$$Y_{ij} = \gamma_{00} + \beta_{1j} X_{ij} + r_{ij} + u_{0j} \tag{30.3}$$

On comparing (30.3) with (30.1), it becomes clear that the only difference between the ordinary regression model and the multilevel model is the additional group-level error term u_{0j}. The basic multilevel model, then, is nothing more than an ordinary regression equation that includes an additional group-level error parameter to capture group-level dependencies.

The addition of the group-level error term explicitly models variation among group means in the data. For example, if the outcome is the mean sentence length given to offenders across federal district courts, the group-level error term allows for mean sentence length to vary by federal district, thus capturing potentially important district-level differences in average punishment severity. These differences are illustrated in Fig. 30.4, where Panel A shows the mean sentence length pooled across a sample of 10 federal districts, and Panel B shows the mean sentence length disaggregated by federal district. The figure indicates that average punishments vary across federal courts. For instance, the mean sentence length in the Northern District of Florida is about twice the average sentence in the District of Delaware. Important differences in variability in punishment also exist across federal districts, with the standard deviation in the Western District of Oklahoma being more than twice that in the Southern District of California. These group-level variations are captured by the incorporation of the group-level error term in the multilevel statistical model, resulting in standard

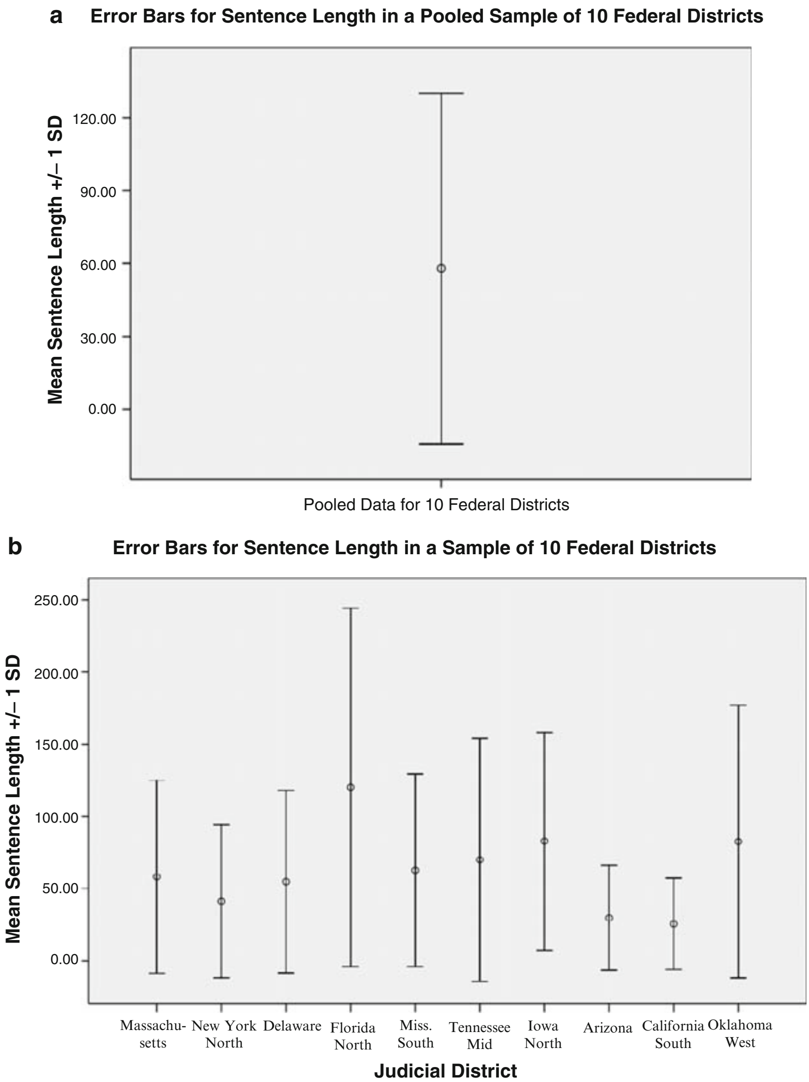


FIGURE 30.4. Variation in sentence lengths across federal courts. Panel (**a**): pooled data for sample of 10 districts. Panel (**b**): disaggregated data for sample of 10 districts.

errors and statistical significance tests that are properly adjusted for the nesting of individual cases within aggregate district court groupings.

Building the Multilevel Model

Despite its conceptual simplicity, multilevel analysis adds a layer of analytical complexity that can quickly become cumbersome when applied to research questions involving multiple predictors across multiple levels of analysis. For this reason, it is essential to thoroughly conduct exploratory data analysis and to carefully build the multilevel model from the ground up. Model misspecifications at one level of analysis can bias parameter estimates across levels of inference. As with any data analysis, multilevel modeling should begin with a careful investigation of the frequency distributions of all individual and ecological variables, as well as detailed investigation of bivariate relationships of interest. The latter should include analysis of cross-level relationships such as unit-specific regressions estimated for all level 2 units. It is also important to investigate the tenability of underlying model assumptions before attempting to estimate full multilevel models (see Raudenbush and Bryk 2002: 255). For instance, assumptions about the normal distribution of error terms apply to each level of analysis and violations of these assumptions can distort statistical inferences. These initial data explorations can provide useful indicators regarding the potential benefits and necessity of using a multilevel model.

Whether or not multilevel analysis is required is both a theoretical and empirical question. First, the research question should always dictate the methodology. Some research questions that involve multiple levels of data may be answerable with simpler and more parsimonious analytical approaches. For instance, one alternative to the multilevel model is to estimate an ordinary regression model along with robust standard errors that are adjusted to account for the clustering of observations within level 2 units. For example, STATA provides a "cluster" command option that adjusts standard errors for residual dependency. This can be a useful approach when the goal is simply to account for clustering, although it does not provide the other advantages of the multilevel model. A second option is to account for group-level variation using a "fixed effects" model that includes dummy variables for each level 2 unit. This is a simple and effective method of accounting for the intraclass correlation due to group dependency, although it precludes examination of between-group differences that are often of theoretical interest.

In general, a minimum number of ecological groupings is recommended before turning to multilevel analysis. This facilitates higher order significance tests and produces more precise estimates of group variance. Raudenbush and Bryk (2002: 267), for example, suggest that for a basic random-intercept model, at least 10 level 2 units are required for each level 2 variable that is included in the model, and for more complicated models more level 2 units are required. Gelman and Hill (2007: 275) have suggested that there is no minimum number of groupings or minimum number of cases per group because in the limiting case (e.g., with only one group or when between-group variances are zero), the multilevel model will reduce to the ordinary regression equation. Although this is technically true, in practice at least a dozen or so clusters is recommended for multilevel modeling in order to satisfy model assumptions regarding a randomly drawn sample and the normal distribution of level 2 error terms. One strength of the multilevel model is that it can handle unbalanced sample designs involving variation in the number of observations per cluster, even when some clusters contain single observations.[8] Ultimately, the researcher must weigh the added complexity introduced by the multilevel model against its analytical advantages for any given research question.

[8] In some research applications, such as couple or twin studies, research designs are routinely constrained to have only 2 observations per grouping. This is not a problem for multilevel modeling. As with ordinary regression, though, small sample sizes (at both level 1 and level 2) can result in low statistical power. Statistical power in multilevel models is more complicated than for ordinary regression models because it is a product of several factors, including the number of clusters, the number of observations per cluster, the strength of the intraclass correlation and the effect sizes for variables in the model. One useful optimal design software program for conducting power analysis with multilevel data is available at: http://sitemaker.umich.edu/group-ased/optimal_design_software.

Before turning to multilevel analysis, it is also useful to begin by testing for the presence of correlated errors. This can be done by estimating an ordinary regression, saving the residuals, and then conducting an analysis of variance to investigate whether or not the residuals are significantly related to group membership. Significant results provide evidence that the ordinary regression assumption of independent errors is violated by the nested structure of the data. Once the decision is made to use a multilevel model, then, there are several types of models that can be estimated. These include (1) unconditional models, (2) random intercept models, (3) random coefficient models, and (4) cross-level interaction models – each adds an additional layer of complexity and provides additional information in the multilevel analysis.

The Unconditional Model. The necessity of multilevel analysis can be further investigated through the *unconditional or null model*. This model is referred to as "unconditional" because it includes no predictors at any level of analysis, so it provides a predicted value for the mean, which is not conditional on any covariates. It is summarized in (30.4):

$$\begin{aligned} &\text{Level 1} \quad Y_{ij} = \beta_{0j} + r_{ij} \\ &\text{Level 2} \quad \beta_{0j} = \gamma_{00} + u_{0j} \end{aligned} \tag{30.4}$$

where Y_{ij} is a continuous outcome for individual i in group j, estimated by the overall intercept β_{0j} plus an individual-level error term, r_{ij}. At level 2 of the model, the intercept β_{0j} is modeled as a product of a level 2 intercept γ_{00} plus a group-level error term, u_{0j}. The unconditional model decomposes the total variance in the outcome into two parts – an individual variance, captured by the individual-level error term, and a group variance, captured by the group-level error term. The unconditional model is therefore useful for investigating the amount of variation that exists within versus between groups. One way to quantify this is to calculate the intraclass correlation coefficient (ICC), which represents the proportion of the total variance that is attributable to between-group differences. The ICC is represented by (30.5):

$$\rho = \frac{\tau_{00}}{(\sigma^2 + \tau_{00})} \tag{30.5}$$

where τ_{00} is the between-group variance estimated by the u_{0j} parameter and σ^2 is the within-group variance estimated by the r_{ij} parameter in (30.4). The intraclass correlation is the ratio of between group variance to total variance in the outcome. Larger ICCs indicate that a greater proportion of the total variance in the outcome is due to between-group differences.[9] It is important to begin any multilevel analysis by estimating the unconditional model. It provides an assessment of whether or not significant between-group variation exists – if it does not, then multilevel analysis is unnecessary – and it serves as a useful baseline model for evaluating explained variance in subsequent model specifications.

Table 30.3 presents the results from an unconditional model examining sentence length for a random sample of federal offenders nested within U.S. district courts. The results are broken into two parts, one for the "fixed effects", which report the unstandardized regression coefficients, and one for the "random effects", which report the variance components for the model. The overall intercept is 52.5 months indicating that the average federal sentence in

[9] It is common in multilevel analysis for between-group variation to represent a relatively small proportion of the total variance, however, as Liska (1990) argues, this does not indicate that between group variation is substantively unimportant.

TABLE 30.3. Unconditional HLM model of federal sentence lengths

Sentence length in months					
Fixed effects	b	S.E.	df	p-value	
Intercept (γ_{00})	52.5	1.8	88	0.00	
Random effects	s^2	S.D.	df	p-value	ρ
Level 1 (r_{ij})	4,630	68			
Level 2 (u_{oj})	267.1	16.3	88	0.00	0.055
Deviance $= 282, 173.7$					
Parameters $= 2$					
$N = 25, 000$					

this sample is just under 5 years. The level 1 variance provides a measure of within-district variation in sentence lengths and the level 2 variance provides an analogous measure for between-district variation. The significance test associated with the level 2 variance component indicates there is significant between-district variation in sentences – sentence lengths vary significantly across federal district courts. Notice that the significance test uses degrees of freedom for the number of level 2 rather than level 1 units; it provides preliminary evidence that districts matter in federal punishment, although as Luke (2004) points out, significance tests for variance components should always be interpreted cautiously.[10]

In order to get a sense of the magnitude of interdistrict variation in punishment, the intraclass correlation coefficient (ICC) can be calculated and the random effects can be assessed in combination with the fixed effects in Table 30.3. The level 2, or between group, variance is $\tau_{00} = 267.1$ and the within-group, or individual variance is $\sigma^2 = 4, 630$. Plugging these values into (30.5) gives an ICC equal to 0.055. This indicates that 5.5% of the total variation in sentence length is attributable to between-district variation in sentencing. Similarly, the standard deviation for the between group variance component can be added and subtracted to the model intercept to provide a range of values for average sentences among districts. Adding and subtracting 16.3 months gives a range between 36.2 and 68.8 months, so the average SENTENCE varies between 3 years and 5 3/4 years for one standard deviation (i.e., about two-thirds) of federal district courts. The significance test, intraclass correlation and range of average sentences all suggest important between-group variation, indicating that multilevel analysis is appropriate in this instance.

THE RANDOM INTERCEPT MODEL. The second type of multilevel model adds predictor variables to the unconditional model and is referred to as a random intercept model because it allows the intercept to take on different values for each level 2 unit in the data. There are three types of random intercept models – models that include only level 1 predictors, models that include only level 2 predictors, and models that include both level 1 and level 2 predictors. In the first model, the focus of the multilevel analysis is on controlling for statistical dependence in clustered observations. In the second, the focus is on estimating variation in group means as a function of group-level predictors, and in the third, the focus in on estimating the joint influence of both level 1 and level 2 predictors. The type of random

[10] Variances are bounded by zero so they are not normally distributed and they are usually expected to take on nonzero values anyway, so it is not always clear what a significant variance means. Although significance tests for variance components can provide a useful starting point, then, they should be used judiciously. It is much more useful to interpret the substantive magnitude of the variance component rather than just its statistical significance.

intercept model will depend on the research question of interest, but it is often useful to begin by estimating the model with only level 1 predictors. This model is presented in (30.6):

$$\begin{aligned} &\text{Level 1} \quad Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \\ &\text{Level 2} \quad \beta_{0j} = \gamma_{00} + u_{0j} \end{aligned} \tag{30.6}$$

where X_{ij} represents an individual-level predictor added to the unconditional model in (30.4). Again, the level 2 equation models the level 1 intercept β_{0j} as a product of both the overall mean intercept, γ_{00}, and a unique level 2 error term, u_{0j}. Substantively, this means that the model intercept is allowed to vary randomly across level 2 units; each level 2 unit in the sample has its own group-specific intercept, just as if separate regressions were estimated for each group in the data.

Table 30.4 presents the results from a model examining the impact of the severity of the offense on the final sentence. In this model, offense severity is centered around its grand mean (see discussion of centering below) and added to the level 1 portion of the model as a predictor of sentence length. β_{1j} in (30.6) represents the effect of offense severity, X_{ij}, on the length of one's sentence in federal court. It is interpreted just as it would be in an ordinary regression model – each one unit increase in offense severity increases one's sentence length by 5.56 months. The average sentence is also allowed to vary by federal district, however. This is reflected by the level 2 variance component u_{0j} in Table 30.4. Both variance components now represent residuals, or left-over variation that is unaccounted for by the model. Notice that the deviance statistic is reduced from the unconditional to the conditional model, indicating increased model fit.[11] To better quantify the model fit, it is often useful to calculate

TABLE 30.4. Random intercept model of federal sentence lengths

Sentence length in months				
Fixed effects	*b*	*S.E.*	*df*	*p-value*
Intercept (γ_{00})	51.0	1.1	88	0.00
Severity (β_1)	5.6	0.2	24,998	0.00
Random effects	s^2	*S.D.*	*df*	*p*-value
Level 1 (r_{ij})	2,228.7	47.2		
Level 2 (u_{oj})	93.2	9.7	88	0.00
Deviance = 263,875.9				
Parameters = 2				
N = 25,000				

[11] The deviance statistic is equal to −2 times the natural log of the likelihood function and serves as a measure of lack of fit between the model and the data – the smaller the deviance the better the model fit. The inclusion of additional predictors will decrease the model deviance, and although the deviance is not directly interpretable, it is useful for comparing alternative model specifications to one another (Luke 2004). The difference in deviance statistics for two models is distributed as a chi-square distribution with degrees of freedom equal to the difference in the number of parameters in the two models. Multilevel models are typically fit with maximum likelihood estimation, but this can be done using either full maximum likelihood (ML) or restricted maximum likelihood (REML). Both estimators will produce identical estimates of the fixed effects, but REML will produce variance estimates that are less biased than ML when the number of level 2 units is relatively small (see Kreft and DeLeeuw 1998: 131–133; Snidjers and Bosker 1999: 88–90). REML is useful for testing two nested models that differ only in their random effects (e.g., an additional random coefficient in the model), but ML must be used to compare models that also differ in their fixed effects (e.g. an additional predictor variable). All example models herein are estimated with REML.

proportionate reduction of error (PRE) measures that approximate R^2 statistics for explained variance at each level of analysis. Equation (30.7) provides the formulas for these calculations:

$$R^2_{\text{Lev1}} = \frac{\sigma^2_{\text{unc}} - \sigma^2_{\text{cond}}}{\sigma^2_{\text{unc}}} \qquad R^2_{\text{Lev2}} = \frac{\tau_{\text{unc}} - \tau_{\text{cond}}}{\tau_{\text{unc}}} \tag{30.7}$$

where explained variation at level 1 is calculated by examining the reduction in level 1 variance relative to the total variance from the unconditional model reported in Table 30.3. The unconditional estimate of level 1 variance was 4,639 and the conditional (i.e., controlling for offense severity) estimate is 2,228.7. This difference (2,401.3) divided by the total unconditional variance (4,630) provides an R^2 estimate of 0.519, so offense severity explains over 50% of the variance in sentence lengths among federal offenders.

The inclusion of level 1 predictors can also explain between-district variation at level 2 of the analysis. This is because there may be important differences in offense severity across districts, with some districts systematically facing more serious crime than others. Explained variation at level 2 is calculated by examining the reduction in level 2 variance from the unconditional to the conditional model. The unconditional estimate for between-district variation was 267.1 and the conditional estimate is 93.2. The difference (173.9) divided by the total (267.1) provides an estimate of explained variation at level 2 equal to 0.651. This indicates that 65% of inter-district variation in sentences is due to the fact that districts vary in the severity of the crimes they face, or 65% of district variation is attributable to *compositional* differences in offense severity.[12]

The random intercept model can be expanded to also include a level 2 predictor as in (30.8):

$$\begin{aligned} &\text{Level 1} \quad Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \\ &\text{Level 2} \quad \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \end{aligned} \tag{30.8}$$

Group mean differences in the intercept, β_{0j}, are now modeled as a product of a group-level predictor, W_j, with γ_{01} representing the effect of the level 2 covariate on the outcome of interest. Level 2 predictors can take several forms including aggregate, structural, or global measures (see Table 30.2). Results for the model including the individual level 1 predictor (offense severity) and the level 2 predictor (Southern location) are presented in Table 30.5. The effect of offense severity remains essentially unchanged, but districts in the South sentence offenders to an additional 7.1 months of incarceration. Although level 1 variables can explain variation at both levels of analysis, level 2 variables can only explain between-group variation at level 2. Accordingly, the level 2 predictor South does not alter the level 1 variance estimate but it does reduce the level 2 variance from 93.2 to 82.4. This is a reduction of 11.6% so Southern location accounts for just under 12% of the residual level 2 variance after controlling for offense severity. Equation (30.8) includes only one level 1 and one level 2 predictor, but the model can be easily expanded to include multiple predictors at both levels of analysis.

[12] These basic formulas for explained variance are simple to apply and often quite useful, but in some circumstances it is possible for the inclusion of additional predictors to result in smaller or even negative values for explained variance (Snijders and Bosker 1999: 99–100). Slightly more complicated alternative formulas are also available that include adjustments for the average number of level 1 units per level 2 unit (see e.g. Luke 2004: 36). Total explained variance at both levels of analysis can be computed using the combined formula: $R^2_{\text{Total}} = \frac{(\sigma^2_{\text{unc}} + \tau_{\text{unc}}) - (\sigma^2_{\text{cond}} + \tau_{\text{cond}})}{\sigma^2_{\text{unc}} + \tau_{\text{unc}}}$

TABLE 30.5. Random intercept model of federal sentence length

Sentence length in months				
Fixed effects	*b*	*S.E.*	*df*	*p-value*
Intercept (γ_{00})	50.9	1.0	87	0.00
South (γ_{01})	7.1	2.3	87	0.00
Severity (β_1)	5.6	0.2	24,997	0.00
Random effects	s^2	*S.D.*	*df*	p-value
Level 1 (r_{ij})	2,228.7	47.2		
Level 2 (u_{oj})	82.4	9.1	87	0.00
Deviance = 263,860.9				
Parameters = 2				
N = 25,000				

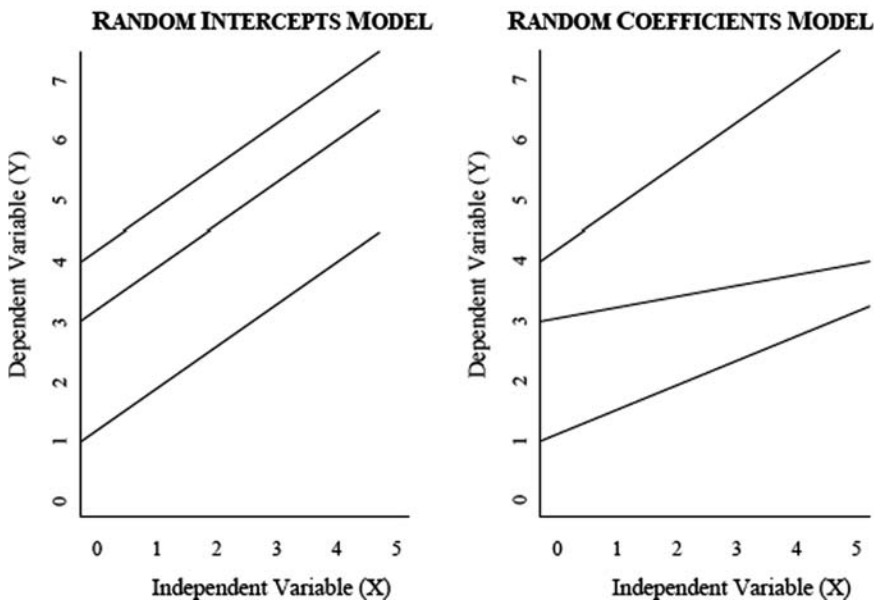


FIGURE 30.5. Comparison of random intercept and random coefficient models.

Although the random intercept model allows the group means to vary as a product of level 2 predictors, it assumes that the effects of the level 1 predictors are uniform across level 2 units. This assumption can be investigated and if it is violated then a random coefficient model may be more appropriate.

THE RANDOM COEFFICIENT MODEL. The random coefficient model builds upon the random intercept model by allowing the effects of individual predictors to also vary randomly across level 2 units. That is, the level 1 slope coefficients are allowed to take on different values in different aggregate groupings. The difference between the random intercept and random coefficient model is graphically depicted in Fig. 30.5, where each line represents the effect of some X on Y for three hypothetical groupings. In the random intercept model, the slopes are constrained to be the same for all three groups, but the intercepts are allowed to be different. In the random coefficient model, both the intercepts and slopes are allowed to differ across the three groups – the effect of X on Y varies by group. Mathematically, the random coefficient model (with a single level 1 predictor) is represented by (30.9):

$$\begin{aligned} &\text{Level 1} \quad Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \\ &\text{Level 2} \quad \beta_{0j} = \gamma_{00} + u_{0j} \\ &\qquad\qquad\; \beta_{1j} = \gamma_{10} + u_{1j} \end{aligned} \tag{30.9}$$

where the key difference from (30.6) is the addition of the new random error term u_{1j} associated with the effect of X_{ij} on Y_{ij}. That is, the β_{1j} slope coefficient is modeled with a random variance component, allowing it to take on different values across level 2 units. For instance, the treatment effect of an after-school delinquency program might vary by school context, being more effective in some schools than others (Gottfredson et al. 2007). The random coefficient model can capture this type of between-group variation in the effect of the independent variable on the outcome of interest.

The decision to specify random coefficients should be based on both theory and empiricism. Regarding federal sentencing data, it might make theoretical sense to investigate variations in the effect of offense severity across courts because some literature suggests perceptions of crime seriousness involve a relative evaluation by court actors (Emerson 1983). Definitions of "serious" crime might be different in different court contexts. To test this proposition, the deviance statistics can be compared for two models, one with offense severity specified as a fixed (i.e., nonvarying) coefficient as reported in Table 30.4 and one with it specified as a random coefficient as in Table 30.6. The deviance for the random intercept model is 263,876 and the deviance for the random coefficient model is 262,530. The difference produces a chi-square statistic of 1,346 with 2 degrees of freedom, which is highly significant.[13] The null hypothesis can therefore be rejected in favor of the random coefficient model.

TABLE 30.6. Random coefficient model of federal sentence length

Sentence length in months				
Fixed effects	*b*	*S.E.*	*df*	*p-value*
Intercept (γ_{00})	49.7	1.0	88	0.00
Severity (β_1)	5.7	0.1	88	0.00
Random effects	s^2	$S.D.$	df	p-value
Level 1 (r_{ij})	2,098.7	45.8		
Level 2 (u_{oj})	78.7	8.9	88	0.00
Severity (u_{1j})	1.4	1.2	88	0.00
Deviance = 262,530.1				
Parameters = 4				
N = 25,000				

[13] The difference in the number of parameters is equal to 2 because the addition of the random coefficient introduces both an additional variance component and an additional covariance component to the model:

$$\text{Var}\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}$$

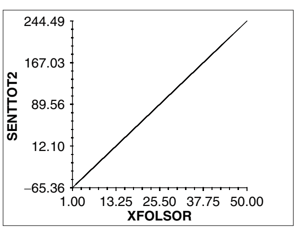
where τ_{11} is the new variance associated with the random coefficient β_{1j}. Because the models only differ in their random components, REML estimation is used for this comparison.

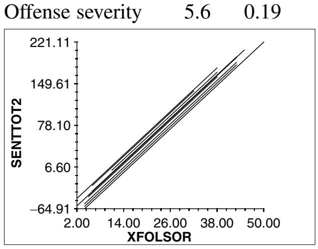
Additional evidence in support of the random coefficient model is provided by the highly significant p-value for the u_{1j} parameter in Table 30.6. This suggests there is significant variation in the effect of offense severity across district courts. To quantify this effect, the standard deviation (S.D. $= 1.2$) for the random effect can be added and subtracted to the coefficient ($b = 5.7$) for offense severity. This suggests that each unit increase in offense severity increases one's sentence length between 4.5 and 6.9 months for one standard deviation (i.e., about two-thirds) of federal district courts. One final diagnostic tool for properly specifying fixed and random coefficients is to compare differences between model-based and robust standard errors.[14] Discrepancies between the two likely indicate model misspecification, such as level 1 coefficients that should be specified as random rather than fixed effects.

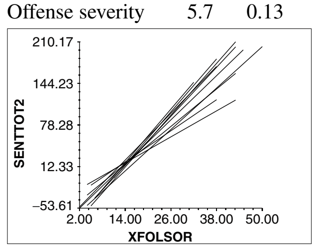
To demonstrate, Table 30.7 provides a comparison of an OLS, random intercept and random coefficient model, along with a pictorial representation of each. As expected, the standard errors in the OLS model are underestimated. The standard error for the model intercept, for instance, increases from 0.30 to 1.10 from the OLS to the random intercept model. Examining the robust standard errors in the random intercept model suggests there may be a problem – the robust standard error for offense severity is more than 6 times as large as its model-based standard error. This is consistent with earlier results that suggested significant variation exists in the effect of offense severity across districts. Allowing for this variation in the random coefficient model produces model-based and robust standard error estimates for offense severity

TABLE 30.7. Comparison of OLS, random intercept and random coefficient models

OLS regression			Random intercept			Random coefficient		
Without robust errors			*Without robust errors*			*Without robust errors*		
	b	S.E.		*b*	S.E.		*b*	S.E.
Intercept	47.9	0.30	Intercept	51.0	1.10	Intercept	49.7	1.02
Offense severity	5.6	0.03	Offense severity	5.6	0.03	Offense severity	5.7	0.13
			With robust errors			*With robust errors*		
				b	S.E.		*b*	S.E.
			Intercept	51.0	1.06	Intercept	49.7	1.01
			Offense severity	5.6	0.19	Offense severity	5.7	0.13
			Deviance = 263,875.9			Deviance = 262,530.1		
			Parameters = 2			Parameters = 4		







[14] Robust standard errors are standard errors that are adjusted to account for possible violations of underlying model assumptions regarding error distributions and covariance structures (see Raudenbush and Bryk 2002: 276). In the case of multilevel models, these violations can lead to misestimated standard errors that result in faulty statistical significance tests. Robust standard errors provide estimates that are relatively insensitive to model misspecifications, but because the calculation of robust standard errors relies on large sample properties, they should only be used when the number of level 2 units is relatively large.

TABLE 30.8. Random coefficient model with level 2 predictor of federal sentence length

Sentence length in months				
Fixed effects	*b*	*S.E.*	*df*	*p-value*
Intercept (γ_{00})	49.6	1.0	87	0.00
South (γ_{01})	3.4	1.4	87	0.02
Severity (β_1)	5.7	0.1	88	0.00
Random effects	s^2	S.D.	df	p-value
Level 1 (r_{ij})	2,098.6	45.8		
Level 2 (u_{oj})	71.6	8.5	87	0.00
Severity (u_{1j})	1.5	1.2	88	0.00

that are identical. Large differences in robust standard errors can serve as a useful diagnostic tool for identifying misspecification in the random effects portion of the multilevel model.

These diagnostic approaches, along with theoretical considerations, should be used to gradually build the random effects portion of the random coefficient model. Ecological predictors can also be included at level 2 of the random coefficient model. Table 30.8 reports the results for the random coefficient model adding Southern location as a level 2 predictor. Notice that the estimated effect of South is less in the random coefficient model in Table 30.8 than it was in the random intercept model in Table 30.5. This highlights the importance of properly specifying the random effects portion of the multilevel model – changes in the random effects at level 1 can alter the estimates for both level 1 and level 2 predictors.

Often times the final multilevel model will include a mixture of fixed and random coefficients, which is why it is sometimes called the "mixed model." Equation (30.10) provides an example of a mixed model with two level 1 predictors and 1 level 2 predictor:

$$
\begin{aligned}
\text{Level 1} \quad & Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + r_{ij} \\
\text{Level 2} \quad & \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \\
& \beta_{1j} = \gamma_{10} + u_{1j} \\
& \beta_{2j} = \gamma_{10}
\end{aligned}
\tag{30.10}
$$

In this mixed model, the effect of the first independent variable X_{1ij} is allowed to have varying effects across level 2 units because its coefficient β_{1j} in level 2 of the model includes the random error term u_{1j}. It is this error variance that allows the effect of X_{1ij} to take on different values for different level 2 units. The effect of the second level 1 predictor X_{2ij}, however, does not include a random error variance. Its effect is therefore constrained to be "fixed" or constant across level 2 units. Although measures of explained variance can be calculated for random coefficient and mixed models, these calculations do not account for the additional variance components introduced by the random effects, so it is advisable to perform these calculations on the random intercept only model (see e.g. Snijders and Bosker 1999: 105).

THE CROSS-LEVEL INTERACTION MODEL. Like ordinary regression models, multilevel models can be further expanded to include interaction terms. These can be incorporated in three basic ways. Individual interactions can be included from cross-product terms for individual-level predictors. For instance, victim race and police officer race might be interacted in a study of police use of force (Lawton 2007). Ecological interactions can also be included using level 2 predictors. Ethnic heterogeneity could be interacted with low socioeconomic conditions at the neighborhood level, for instance, in a study of risk of victimization

(Miethe and McDowall 1993).[15] Finally, cross-level interactions can be included that specify cross-product terms across-levels of analysis. For instance, the effects of parental monitoring on problem behavior at the individual level might be expected to vary among neighborhoods with different levels of collective efficacy (Rankin and Quane 2002). This type of interaction is unique to multilevel analysis so it deserves additional explanation. Equation (30.11) specifies a cross-level interaction model with 1 individual predictor, 1 ecological predictor and the cross level interaction between them:

$$\begin{aligned} &\text{Level 1} \quad Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + r_{ij} \\ &\text{Level 2} \quad \beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j} \\ &\qquad\qquad\; \beta_{1j} = \gamma_{10} + \gamma_{11} W_j + u_{1j} \end{aligned} \tag{30.11}$$

This model adds the level 2 predictor W_j to the level 2 equation for β_{1j}, so W_j is now being used to explain variation in the effect of β_{1j} across level 2 units, with the new parameter γ_{11} representing the cross-level interaction between X_{1ij} and W_j. Cross-level interactions are useful for answering questions about *why* individual effects vary across level 2 units; they explicitly model variation in level 1 random coefficients as a product of level 2 group characteristics.[16] Table 30.9 provides results from a cross-level interaction model examining the conditioning effects of Southern court location on the individual effect of offense severity for federal sentence lengths. The positive interaction effect indicates that offense severity has a stronger effect on sentence length in Southern districts than it does in non-Southern districts. As with other multilevel models, cross-level interactions models can easily be extended to the case of multiple predictors at both the individual and group levels of analysis, although care should be taken when including multiple interactions in the same model.

TABLE 30.9. Cross-level interaction model of federal sentence length

Sentence length in months					
Fixed effects	*b*	*S.E.*	*df*	*T-ratio*	*p-value*
Intercept (γ_{00})	49.6	1.0	87	49.14	0.00
South (γ_{01})	6.3	2.0	87	3.37	0.02
Severity (β_1)	5.7	0.1	88	42.45	0.00
South*severity (γ_{11})	0.6	0.3	87	2.07	0.04
Random effects	s^2	S.D.	df	χ^2	p-value
Level 1 (r_{ij})	2,098.6	45.8			
Level 2 (u_{oj})	70.2	8.4	87	1130.14	0.00
Severity (u_{1j})	1.4	1.2	88	1644.40	0.00
Deviance = 262,519.3					
Parameters = 4					
N = 25,000					

[15] Depending on the statistical program used, these interactions may or may not be able to be created in the multilevel interface. With HLM, both individual interactions and ecological interactions must be created and all centering adjustments must be made before importing them into the HLM program.

[16] Although conceptually the goal of cross-level interactions is usually to explain significant variation in the effects of level 1 random coefficients across level 2 units, there are instances when theory may dictate examining cross-level interactions for fixed coefficients at level 1 as well. Significant cross-level interactions may emerge involving fixed level 1 coefficients because the significance tests for the cross-level interactions are more powerful than the significance tests produced for random coefficient variance components (Snijders and Bosker 1999: 74–75).

ADDITIONAL CONSIDERATIONS

The preceding examples offer only a rudimentary introduction to the full gamut of multilevel modeling applications, but they provide a basic foundation for doing more complex multilevel analysis. The multilevel model can be further adapted to account for additional data complexities that commonly arise in criminological research, including centering conventions, nonlinear dependent variables, and additional levels of analysis. These issues are briefly highlighted below although interested readers should consult comprehensive treatments available elsewhere (e.g., Raudenbush and Bryk 2002; Luke 2004; Goldstein 1995; Snidjers and Bosker 1999; Kreft and de Leeuw 1998; Gelman and Hill 2007).

Centering in Multilevel Analysis

In multilevel models, the centering of variables takes on special importance. Centering, or reparameterization, involves simple linear transformations of the predictor variables by subtracting a constant such as the mean of X or W. Centering in the multilevel framework is no different than in ordinary multiple regression, but it offers important analytical advantages, making model intercepts more interpretable, making main effects more meaningful when interactions are included, reducing collinearity associated with polynomials and interactions, facilitating model convergence in nonlinear models, and simplifying graphical displays of output. Estimates of variance components may also be affected by the centering convention because random coefficients often involve heteroskedastic error variances that depend on the value of X at which they are evaluated (Hox 2002).

In general, three main centering options are available: no centering, grand-mean centering and group-mean centering. No centering leaves the variable untransformed in its original metric. Although this can be a reasonable approach depending on how the variables are measured, it is usually advisable to employ a centering convention in multilevel analyses for the reasons stated above. The simplest centering convention is grand-mean centering, which involves subtracting the overall mean, or the pooled average, from each observation in the data. The subtracted mean, then, becomes the new zero point so that positive values represent scores above the mean and negative values represent scores below the mean. Grand mean centering is represented as $(X_{ij} - \overline{X}_{..})$, where X_{ij} is the value of X for individual i in group j and $\overline{X}_{..}$ is the grand mean pooled across all observations in the data. Grand mean centering is often useful and rarely detrimental so it offers a good standard centering convention. It only affects the parameter estimates for the model intercept, making the value of the intercept equal to the predicted value of Y when all variables are set to their means. This allows the intercept in a grand-mean centered model to be interpreted as the expected value for the "average" observation in the data.

The alternative to grand mean centering is group mean centering, represented as $(X_{ij} - \overline{X}_{.j})$, where X_{ij} is still the value of X for individual i in group, j but $\overline{X}_{.j}$ is now the group-specific mean, so individuals in different level 2 groups have different values of $\overline{X}_{.j}$ subtracted from their scores. Group-mean centering is more complicated than grand-mean centering because it fundamentally alters the meaning and interpretation of both the parameter estimates and the variance components in the multilevel model. Group-mean centering can be a useful diagnostic tool and it is appropriate for many research questions, but in general it should be used selectively. Luke (2004: 52), for instance, recommended that "one should use group-mean centering only if there are strong theoretical reasons to do so."[17]

In general, centering is always a good idea when a variable has a nonmeaningful zero point. For example, it would make little sense to include the UCR crime rate as a predictor variable without first centering it. Otherwise, the model intercept would represent the predicted value of Y when the crime rate was equal to 0, which is clearly unrealistic. Even when variables do have meaningful zero points, it is often useful to center them. For instance, often times it is even useful to center dummy variables. Adjusting for the grand mean essentially removes the influence of the dummy variable so that the model intercept represents the expected value of Y for the "average" of that variable rather than for the reference category. Similar centering rules apply for ecological variables as for individual level variables, but the important point is that centering decisions should be made *a priori* based on theoretical considerations regarding the desired meaning of model parameters. A number of more detailed treatments offer further detail on the merits and demerits of grand-mean and group-mean centering conventions for multilevel analysis (e.g., Kreft 1995; Kreft et al. 1995; Longford 1989; Raudenbush 1989; Paccagnella 2006).

Generalized Multilevel Models

The examples up to this point all assume a normally distributed continuous dependent variable. Often times, however, criminological research questions involve nonlinear or discrete outcomes, such as binary, count, ordinal, or multinomial variables. When this is the case, the multilevel model must be adapted by transforming the dependent variable. For example, dichotomous dependent variables are common in research on crime and justice; whether or not an offender commits a crime, the police make an arrest, or a judge sentences to incarceration, all involve binary outcomes (e.g. Eitle et al. 2005; Griffin and Armstrong 2003; Johnson 2006). In these cases, the discrete dependent variable often violates assumptions of the general linear model regarding linearity, normality, and homoskedasticity of level 1 errors (Raudenbush and Bryk 2002). Moreover, because the outcome is bound by 0 and 1, the fitted linear model is likely to produce nonsensical and out of range predictions.

None of these issues are unique to multilevel analysis and the same adjustments used in ordinary regression can be applied to the multilevel model, although some important new issues arise in the multilevel context. Collectively, these types of models are labeled generalized hierarchical linear models (GHLM) or just generalized multilevel models, because they provide flexible generalizations of the ordinary linear model. The basic structure of the multilevel model remains the same, but the sampling distribution changes. For illustrative purposes, the case of multilevel logistic regression with a dichotomous outcome is illustrated. Equation

[17] Some exceptions to this general rule include growth curve modeling with longitudinal data, where the focus is often on separating within and between group regression effects, or research questions involving "frog pond" effects, where the theoretical interest is on individual adaptation to one's specific environment rather than the average effects of individual predictors on the outcome of interest.

(30.12) provides the formula for the unconditional two-level multinomial logistic model using the binomial sampling distribution and the logit link function:[18]

$$\begin{array}{lrl} \text{Logit Link Function} & \eta_{ij} & = \ln\left(\dfrac{p}{1-p}\right) \\ \text{Level 1} & \eta_{ij} & = \beta_{0j} \\ \text{Level 2} & \beta_{0j} & = \gamma_{00} + u_{0j} \end{array} \tag{30.12}$$

In this formulation, p is the probability of the event occurring and $(1 - p)$ is the probability of the event not occurring. p over $(1 - p)$, then, represents the *odds* of the event and taking the natural log provides the *log odds*. The dependent variable for the dichotomous outcome is therefore the log of the odds of success for individual i in group j, represented by η_{ij}. The multinomial logistic model is probabilistic, capturing the likelihood that the outcome occurs. Whereas the original binary outcome was constrained to be 0 or 1, p is allowed to vary in the interval 0 to 1, and η_{ij} can take on any real value. In this way, the logistic link function transforms the discrete outcome into a continuous range of values. The level 2 model is identical to that for the continuous outcome presented in (30.4), but γ_{00} now represents the average log odds of the event occurring across all level 2 units. Equation (30.13) provides the random coefficient extension of the multilevel logistic model with one random level 1 coefficient and one level 2 predictor:

$$\begin{array}{lrl} \text{Level 1} & \eta_{ij} & = \beta_{0j} + \beta_{1j} X_{1ij} \\ \text{Level 2} & \beta_{0j} & = \gamma_{00} + \gamma_{01} W_j + u_{0j} \\ & \beta_{1j} & = \gamma_{10} + u_{1j} \end{array} \tag{30.13}$$

where η_{ij} still represents the log of the odds of success and all the other parameters are the same as previously described.

Notice that in both (30.12) and (30.13), there is no level 1 variance component included in the multilevel logistic model. This is because the level 1 variance is heteroskedastic and completely determined by the value of p, it is therefore unidentified and not included in the model. This means that the standard formulas for the intraclass correlation and explained variance at level 1 cannot be directly applied to the case of a binary dependent variable.[19] Also, most software packages do not provide deviance statistics for nonlinear multilevel models. This is because generalized linear models typically rely on "penalized quasi likelihood" (PQL), rather than full or restricted maximum likelihood. This involves a double-iterative process that provides only a rough approximation to the likelihood function on which the deviance is based. In most cases, this means that other methods, such as theory, significance tests for

[18] The "link function" can be thought of as a mathematical transformation that allows the nonnormal dependent variable to be linearly predicted by the explanatory variables in the model.

[19] The level 1 variance in the case of a logistic model is equal to $p(1 - p)$ where p is the predicted probability for the level 1 model. The level 1 variance therefore varies as a direct product of the value of p at which the model is evaluated. Although multilevel logistic models do not include a level 1 variance term, some alternatives approaches are available for estimating intraclass correlations. For example, Snijders and Bosker, (1999: Chap. 14) discuss reconceptualizing the level 1 model as a latent variable $Z_{ij} = \eta_{ij} + r_{ij}$, in which the level 1 error term is assumed to have a standard logistic distribution with a mean of 0 and variance of $\pi^2/3$. In that case, the intraclass correlation can be calculated as $\rho = \tau_{00}/(\tau_{00} + \pi^2/3)$. This formulation requires the use of the logit link function and relies on the assumption that the level 1 variance follows the logistic distribution. Alternative formulations have also been discussed for the probit link function using the normal distribution (see e.g. Gelman and Hill 2007: 118).

variance components, and robust standard error comparisons must be relied on to properly specify random coefficients in level 1 of the multilevel logistic model.[20]

A third complication involving multilevel models with nonlinear link functions is that two sets of results are produced, one labeled "unit-specific" results and one labeled "population-average" results. Unit-specific results are estimated holding constant the random effects in the model, whereas population-average results are averaged across all level 2 random effects (see Raudenbush and Bryk 2002: 301). This means that unit-specific estimates model the dependent variable conditional on the random effects in the model, which provide estimates of how the level 1 and level 2 variables affect outcomes *within* the level 2 units. Population-average estimates, provide the marginal expectation of the outcome averaged across the entire population of level 2 units. If you wanted to know how much an after-school program reduces delinquency for one student compared to another in the same school, then the unit-specific estimate would be appropriate. If you wanted to summarize the average effect of the after-school program on delinquency across all schools, then the population-average estimate would be preferred. In short, which results to report depends on the research question at hand.[21] For example, work on racial disparity in sentencing typically reports unit-specific estimates because the focus is on the effect of an offender's race relative to other offenders sentenced in the same court (e.g., Ulmer and Johnson 2004). Recent work integrating routine activities and social disorganization theory, on the other hand, reports population average estimates because in the words of the authors "our research questions concern aggregate rates of delinquency and unstructured socializing" among all schools (Osgood and Anderson 2004: 534).

Table 30.10 reports the unit-specific results with robust standard errors for a random coefficient model examining the likelihood of imprisonment in federal court. The level 1 predictor is the severity of the offense and the level 2 predictor is Southern location. Offense severity exerts a strong positive effect on the probability of incarceration. The coefficient of 0.26 represents the change in the log odds of imprisonment for a one-unit increase in severity. To make this more interpretable, it is useful to transform the raw coefficient into an odds ratio.

TABLE 30.10. Multilevel logistic model of federal incarceration

Prison vs. no prison (unit-specific model with robust standard errors)					
Fixed effects	*b*	*S.E.*	*df*	*p-value*	*Odds ratio*
Intercept (γ_{00})	2.80	0.08	87	0.00	
South (γ_{01})	0.07	0.11	87	0.53	1.07
Severity (β_1)	0.26	0.01	88	0.00	1.29
Random effects	s^2	S.D.	df	p-value	
Level 2 (u_{0j})	0.41	0.64	87	0.00	
Severity (u_{1j})	0.004	0.06	88	0.00	

[20] PQL estimates are usually sufficient, but tests for random effects based on the PQL likelihood function in models with discrete outcomes may be unreliable, especially for small samples. Alternative full maximum estimators, such as Laplace estimation, are available in some software packages and can be used to test for random effects using the deviance, but this can be computationally intensive.

[21] These estimates are often similar, but their differences will widen as between-group variance increases and the probability of the outcome becomes farther away from.0.50 (Raudenbush and Bryk, 2002: 302). In the case of continuous dependent variables, the unit-specific and population estimates are identical so this distinction only arises in the case of nonlinear dependent variables.

Because the left-hand side of (30.13) represents the log of the odds, we obtain the odds by taking the antilog, in this case $e^{0.256} = 1.29$. For each unit increase in the severity of the crime committed, the odds of incarceration increases by a factor of 0.29 or 29%.[22] The coefficient for South in this model is not statistically significant, suggesting there is no statistical evidence that offenders are more likely to be incarcerated in Southern districts. Turning to the random effects, the level 2 intercept indicates that significant interdistrict variation in incarceration remains after controlling for severity and Southern location, and that significant variance exists in the effect of offense severity across districts. Adding the standard deviation to the fixed effect for severity provides a range of coefficients between 0.20 and 0.32. Transformed into odds ratios, this means that the effect of offense severity varies between 1.22 and 1.38, so offense severity increases the odds of incarceration between 22% and 38% across one standard deviation (i.e., about two-thirds) of federal districts.

As with linear multilevel models, generalized multilevel models can be easily extended to the case of multiple predictors at both levels of analysis. In general, similar transformations can be applied for multilevel Poisson, binomial, ordinal, and multinomial models by simply applying different link functions to different sampling distributions (see e.g., Raudenbush and Bryk 2002: Chap. 10; Luke 2004: 53–62).[23] In this way, the basic linear multilevel model can be easily generalized to address a variety of criminological research questions involving different types of discrete dependent variables.

Three-Level Multilevel Models

The basic two-level multilevel linear and generalized models can also be extended to incorporate more complicated data structures that span three or more levels of analysis.[24] The basic logic of the multilevel model is the same, but additional error variances are added for each additional level of analysis. The three-level unconditional model for a linear dependent variable is presented in (30.14):

$$\begin{array}{ll} \text{Level 1} & Y_{ijk} = \pi_{0jk} + e_{ijk} \\ \text{Level 2} & \pi_{0jk} = +\beta_{00k} + r_{0jk} \\ \text{Level 3} & \beta_{00k} = \gamma_{000} + u_{00k} \end{array} \tag{30.14}$$

The i subscript indexes level 1 (e.g., students), the j subscript indexes level 2 (e.g., classrooms), and the k subscript indexes level 3 (e.g. schools). Now level 1 coefficients are represented with π's, level 2 coefficients with β's and level 3 coefficients with γ's, but the three-level structure is purely notational convenience, so it can be simplified through substitution to produce the equivalent, but simpler, combined model in (30.15):

$$Y_{ijk} = \gamma_{000} + e_{ijk} + r_{0jk} + u_{00k} \tag{30.15}$$

[22] The individual probability of incarceration for individual i in court j can be calculated directly using the formula: $p_{ij} = \frac{e^{\gamma_{00}+\gamma_{01}W_j+\gamma_{10}X_{ij}}}{(1+e^{\gamma_{00}+\gamma_{01}W_j+\gamma_{10}X_{ij}})}$, so with grand-mean centering the mean probability of incarceration is $\bar{p}_{ij} = \frac{e^{\gamma_{00}}}{(1+e^{\gamma_{00}})}$.

[23] Some important differences emerge in these other contexts, for example, overdispersion frequently occurs in Poison models for count data, so it is common to incorporate an additional overdispersion parameter in the level 1 model for this type of generalized linear model (see Raudenbush and Bryk 2002: 295; Gelman and Hill 2007: 114).

[24] Some software packages, such as, HLM are currently limited to three levels of analysis, but other programs (e.g., WLwiN) can analyze up to 10 separate levels of analysis.

TABLE 30.11. Three-level unconditional model of federal sentence length

Sentence length in months					
Fixed effects	*b*	*S.E.*	*df*	*p-value*	
Intercept (γ_{000})	52.5	3.2	10	0.00	
Random effects	s^2	S.D.	df	p-value	ρ
Level 1 (e_{ijk})	4,630.1	68.0			
Level 2 ($r_{0\text{jk}}$)	172.5	15.2	78	0.00	0.035
Level 3 ($u_{00\text{k}}$)	85.2	9.2	10	0.00	0.017

Equations (30.14) and (30.15) are substantively identical and it becomes clear in the combined model that the outcome Y_{ijk} is modeled as a simple product of an overall intercept γ_{000} plus three different error terms, one for each level of analysis. As in the case of the two-level unconditional model, the three-level model parcels the variation in the outcome across levels of analysis. Similar estimates can therefore be calculated for intraclass correlation coefficients, but in the case of the three-level model, there are separate ρ coefficients for level 2 and level 3 of the analysis.[25]

In the federal court system, cases are nested within district courts, but district courts are also nested within circuit courts, which serve as courts of appeal and play an important role in establishing federal case law. Table 30.11 provides the results from a three-level unconditional model examining federal sentence lengths for the same random sample of 25,000 cases, nested within 89 federal districts, and within 11 federal circuits. The level 2 and level 3 variance components are highly significant, indicating that federal sentences vary significantly across both district and circuit courts. The intraclass correlation coefficients suggest that about 3.5% of the total variation sentencing is between federal districts with another 1.7% between circuit courts. Notice that some of the between-district court variation from the two-level model in Table 30.3 is now being accounted for by level 3 of the analysis.

As with the two-level model, predictors can be added at each level of analysis. That is, individual predictors can be added at level 1, district court predictors can be added at level 2, and circuit court predictors can be added at level 3. Similar steps can then be taken to identify random coefficients as with the two-level model, but care should be exercised in this process because error structures for three-level models can quickly become complicated. This is because Level 1 variables can be specified as random coefficients at *both* level 2 *and* level 3 of the analysis. Moreover, Level 2 coefficients can also be specified as random effects at level 3 of the analysis. Cross-level interactions can occur between levels 1 and 2, levels 1 and 3, or levels 2 and 3. The various possible model specifications can quickly become unwieldy, so it is particularly important in three-level models to exercise care in first identifying the hypothesized effects of interest and then properly specifying the model to capture them.

[25] The formula for the level 2 intraclass correlation is $\rho_{\text{Level 2}} = \tau_\pi/(\sigma^2 + \tau_\pi + \tau_\beta)$, where σ^2 is the level 1 variance, τ_π is the level 2 variance, and τ_β is the level 3 variance. The formula for the level 3 intraclass correlation is $\rho_{\text{Level 3}} = \tau_\beta/(\sigma^2 + \tau_\pi + \tau_\beta)$ (see Raudenbush and Bryk 2002: 230).

Equation (30.16) provides an example of a basic three-level mixed model with one level 1 predictor, Z_{ijk}, specified as randomly varying a cross both level 2 and level 3, one level 2 predictor, X_{jk}, fixed at level 3, and no level 3 predictors:

$$\begin{aligned}
&\text{Level 1} \quad Y_{ijk} = \pi_{0jk} + \pi_{1jk} Z_{ijk} + e_{ijk} \\
&\text{Level 2} \quad \pi_{0jk} = \beta_{00k} + \beta_{01k} X_{jk} + r_{0jk} \\
&\qquad\qquad \pi_{1jk} = \beta_{10k} + r_{1jk} \\
&\text{Level 3} \quad \beta_{00k} = \gamma_{000} + u_{00k} \\
&\qquad\qquad \beta_{01k} = \gamma_{010} \\
&\qquad\qquad \beta_{10k} = \gamma_{100} + u_{10k}
\end{aligned} \tag{30.16}$$

The subscripts and multiple levels can easily become confusing, so it is often useful to examine the combined model, substituting levels 2 and 3 into the level 1 equation. Equation (30.17) provides this reformulation with the fixed effects, or regression coefficients, isolated with parentheses and the random effects, or error variances, isolated with brackets:

$$Y_{ijk} = (\gamma_{000} + \gamma_{100} Z_{ijk} + \gamma_{010} X_{jk}) + [e_{ijk} + r_{0jk} + u_{00k} + r_{1jk} Z_{ijk} + u_{10k} Z_{ijk} +] \tag{30.17}$$

γ_{000} is the overall model intercept, and γ_{100} and γ_{010} are the regression effects for the level 1 and level 2 predictors. As in the unconditional model, e_{ijk}, r_{0jk}, and u_{00k} are the level 1, 2, and 3 error variances, and the new error terms, $r_{1jk} Z_{ijk}$ and $u_{10k} Z_{ijk}$, indicate that the effect of the level 1 variable, Z_{ijk}, is allowed to vary across both level 2 and level 3 units.

Estimating this model with data on federal sentence lengths produces the output in Table 30.12. These results report model-based rather than robust standard errors because the highest level of analysis includes only 11 circuit courts. The effect of offense severity is essentially the same, increasing sentence length by about 5.7 months, but the effect for Southern location has been attenuated and is now only marginally significant. This likely reflects the fact that some of the district variation is now being accounted for by the circuit level of analysis. The random effects in Table 30.11 support this interpretation. The level 2 variance component is smaller than it was in the two-level model reported in Table 30.8. Notice also that there are two variance components associated with offense severity because its effect is allowed to vary both across district and circuit courts. The magnitude of these variance components indicates there is more between-district than between-circuit variation in the effect of offense severity, but both are highly significant. Although conceptually the

TABLE 30.12. Three-level mixed model of federal sentence length

Sentence length in months				
Fixed effects	*b*	*S.E.*	*df*	*p-value*
Intercept (γ_{000})	48.6	1.6	10	0.00
South (β_{01k})	2.8	1.7	87	0.10
Severity (π_{1jk})	5.7	0.2	10	0.00
Random effects	s^2	S.D.	df	p-value
Level 1 (e_{ijk})	2,098.6	45.8		
Level 2 (r_{0jk})	53.6	7.3	77	0.00
Severity (r_{1jk})	1.1	1.1	78	0.00
Level 3 (u_{00k})	17.6	4.2	10	0.00
Severity (u_{10k})	0.3	0.5	10	0.00

three-level multilevel model represents a straightforward extension of the two-level model, in practice care needs to be exercised to avoid exploding complexity (for recent examples using 3 level models see Duncan et al. 2003; Johnson 2006; Wright et al. 2007).

SUMMARY AND CONCLUSIONS

Multilevel models represent an increasingly popular analytical approach in the field of criminology. According to a recent analysis by Kleck et al. (2006), between 5% and 6% of empirical research papers in top criminology journals utilize multilevel modeling, but given the prevalence of multilevel research questions, their use is likely to continue to gain prominence in the field. Because multilevel models provide a sophisticated approach for integrating multiple levels of analysis, they represent an important opportunity to expand theoretical and empirical discourse across a variety of criminological domains. Multilevel models have already been used to study a rich diversity of topics, from examinations of self-control (Hay and Forrest 2006; Doherty 2006; Wright and Beaver 2005) and strain theory (Slocum et al. 2005) to life course perspectives (Horney et al. 1995; Sampson et al. 2006) and analyses of violent specialization (Osgood and Schreck 2007) – from crime victimization (Xie and McDowall 2008; Wilcox et al. 2007), policing (Rosenfeld et al. 2007; Warner 2007), and punishment outcomes (Kleck et al. 2005; Bontrager et al. 2005; Johnson 2005; 2006) to postrelease recidivism (Kubrin and Stewart 2005; Chiricos et al. 2007; Mears et al. 2008) and program evaluations (Gottfredson et al. 2007; Esbensen et al. 2001) – across a broad range of criminological topic areas, multilevel models have proven to be invaluable tools.

Despite their many applications, though, the old adage that "A little bit of knowledge can be a dangerous thing" applies directly to multilevel modeling. Modern software packages make estimating multilevel models relatively simple, but the fully specified multilevel model often contains complicated error structures that can easily be misspecified. Moreover, these complexities can sometimes result in instability in parameter estimates. This is particularly the case for ecological predictors and for three-level and generalized linear models. For instance, it is common for ecological predictors to have shared variance (Land et al. 1990), so inclusion or elimination of one predictor can often affect the estimates for other predictors in the model. It is therefore essential that the final model be carefully constructed from the ground up, performing model diagnostics to test for misspecification, investigating problematic collinearity and examining alternative models to ensure that the final estimates are robust to minor alterations in model specification.

Although this chapter provides a basic overview of multilevel models, it is important to note that it does not cover many of their advanced applications such as longitudinal data analysis, growth-curve modeling, time series data, latent variable analysis, meta-analytical techniques, or analysis of cross-classified data. Beyond situations where individuals are influenced by social contexts, multilevel data commonly characterizes these and many other criminological enterprises. As a discipline, we are just beginning to incorporate the full range of applications for multilevel statistical models in the study of crime and punishment. The goals of this chapter were simply to introduce the reader to the basic multilevel model, to emphasize the ways in which it is similar to and different from the ordinary regression model, to provide some brief examples of different types of multilevel models, and to demonstrate how they can be estimated within the context of jurisdictional variations in federal criminal punishments across court contexts.

Acknowledgment The author would like to thank and acknowledge the very valuable comments provided by D. Wayne Osgood on an earlier draft of this chapter.

REFERENCES

Beaver KM, Wright JP, Maume MO (2008) The effect of school classroom characteristics on low self-control: a multilevel analysis. J Crim Justice 36:174–181

Bernburg JG, Thorlindsson T (2007) Community structure and adolescent delinquency in Iceland: a contextual analysis. Criminology 45(2):415–444

Blalock HM (1984) Contextual-effects models: theoretical and methodological issues. Annu Rev Sociol 10:353–372

Blau PM (1960) Structural effects. Am Sociol Rev 25:178–193

Blumer H (1958) Race prejudice as a sense of group position. Pacific Sociol Rev 1:3–7

Bontrager S, Bales W, Chiricos T (2005) Race, ethnicity, threat and the labeling of convicted felons. Criminology 43(3):589–622

Britt CL (2000) Social context and racial disparities in punishment decisions. Justice Q 17(4):707–732

Camp S, Gaes G, Langan N, Saylor W (2003) The influence of prisons on inmate misconduct: a multilevel investigation. Justice Q 20(3):501–533

Chiricos T, Barrick K, Bales W, Bontrager S (2007) The labeling of convicted felons and its consequences for recidivism. Criminology 45(3):547–581

Diez Roux AV (2002) A glossary for multilevel analysis. J Epidemiol Community Health 56:588–594

Doherty EE (2006) Self-control, social bonds, and desistance: a test of life-course interdependence. Criminology 44(4):807–833

Duncan TE, Duncan SC, Okut H, Strycker LA, Hix-Small H (2003) A multilevel contextual model of neighborhood collective efficacy. Am J Community Psychol 32(3):245–252

Eitle D, Stolzenberg L, D'Alessio SJ (2005) Police organizational factors, the racial composition of the police, and the probability of arrest. Justice Q 22(1):30–57

Emerson RM (1983) Holistic effects in social control decision-making. Law Soc Rev 17(3):425–456. oHol

Finn-Aage E, Wayne Osgood D, Taylor TJ, Peterson D (2001) How great is G.R.E.A.T.? Results from a longitudinal quasi-experimental design. Criminol Public Policy 1(1):87–118

Firebaugh G (1978) A rule for inferring individual-level relationships from aggregate data. Am Sociol Rev 43: 557–572

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

Gillespie W (2005) A multilevel model of drug abuse inside prison. Prison J 85(2):223–246

Goldstein H, Yang M, Omar R., Turner R, Thompson S (2000) Meta-analysis using multilevel modelswith an application to the study of class size effects. Appl Stat 49:399–412

Goldstein H (1995) Multilevel statistical models. 2nd edn. Wiley, New York

Gottfredson DC, Cross A, Soule DA (2007) Distinguishing characteristics of effective and ineffective after-school programs to prevent delinquency and victimization. Criminol Public Policy 6(2):289–318

Griffin ML, Armstrong GS (2003) The effect of local life circumstances on female probationers' offending. Justice Q 20(2):213–239

Griffin T, Wooldredge J (2006) Sex-based disparities in felony dispositions before versus after sentencing reform in Ohio. Criminology 44(4): 893–923

Hay C, Forrest W (2006) The development of self-control: examining self-control theory's stability thesis. Criminology 44(4):739–774

Horney J, Osgood DW, Marshall IH (1995) Criminal careers in the short-term: intra-individual variability in crime and its relation to local life circumstances. Am Sociol Rev 60(4):655–673

Hox J (2002) Multilevel analysis: techniques and applications. Lawrence Erlbaum Associates, Inc. Publishers, Mahwah, NJ

Huebner BM (2003) Administrative determinants of inmate violence: a multilevel analysis. J Crim Justice 31(2): 107–117

Johnson BD (2005) Contextual disparities in guidelines departures: courtroom social contexts, guidelines compliance, and extralegal disparities in criminal sentencing. Criminology 43(3):761–796

Johnson BD (2006) The multilevel context of criminal sentencing: integrating judge- and county-level influences. Criminology 44(2):259–298

Johnson BD, Ulmer JT, Kramer JH (2008) The social context of guidelines circumvention: the case of federal district courts. Criminology 46(3):737–783

Kleck G, Sever B, Li S, Gertz M (2005) The missing link in general deterrence research. Criminology 43(3):623–660

Kleck G, Tark J, Bellows JJ (2006) What methods are most frequently used in research in criminology and criminal justice? J Crim Justice 36(2):147–152

Kreager DA (2007) When it's good to be "bad": violence and adolescent peer acceptance. Criminology 45(4):893–923

Kreft IGG (1995) The effects of centering in multilevel analysis: is the public school the loser or the winner? A new analysis of an old question. Multilevel Modeling Newsletter 7:5–8

Kreft IGG, De Leeuw J, Aiken LS (1995) The effect of different forms of centering in hierarchical linear models. Multivariate Behav Res 30(1):1–21

Kreft IGG, De Leeuw J (1998) Introducing multilevel modeling. Sage Publications, Thousand Oaks

Kubrin CE, Stewart EA (2006) Predicting who reoffends: the neglected role of neighborhood context in recidivism studies. Criminology 44(1):165–197

Land KC, McCall PL, Cohen LE (1990) Structural covariates of homicide rates: are there any invariances across time and social space? Am Sociol Rev 95(4):922–963

Lauritsen JL, Schaum RJ (2004) The social ecology of violence against women. Criminology, 42(2):323–357

Lawton BA (2007) Levels of nonlethal force: an examination of individual, situational and contextual factors. J Res Crime and Delinq 44(2):163–184

Lee MS, Ulmer JT (2000) Fear of crime among Korean Americans in Chicago communities. Criminology 38(4):1173–1206

Liska AE (1990) The Significance of aggregate dependent variables and contextual independent variables for linking macro and micro theories. Soc Psych Q 53:292–301

Liska AE (1992) Social threat and social control. SUNY Press, Albany

Longford NT (1989). To center or not to center. Multilevel Modelling Newsletter 1(2):7–11

Luke DA (2004) Multilevel modeling. Sage Publications, Thousand Oaks

Mears DP, Wang X, Hay C, Bales WD (2008) Social ecology and recidivism: implications for prisoner reentry. Criminology 46(2):301–340

Miethe TD, McDowall D (1993) Contextual effects in models of criminal victimization. Soc Forces 71(3):741–759

Mill, John S. (1843). A system of logic, ratiocinative, and inductive. Longmans, Green, Reader, and Dyer. (8th edn, 1872) or "Harrison & Co." for 1st edn

Nieuwbeerta P, McCall PL, Elffers H, Wittebrood K (2008) Neighborhood characteristics and individual homicide risks: effects of social cohesion, confidence, in the police, and socioeconomic disadvantage. Homicide Studies 12(1):90–116

Osgood DW, Anderson AL (2004) Unstructured socializing and rates of delinquency. Criminology 42(3):519–550

Osgood DW, Schreck CJ (2007) A new method for studying the extent, stability, and predictors of individual specialization in violence. Criminology 45(2):273–312

Paccagnella O (2006) Centering or not centering in multilevel models? The role of the group mean and the assessment of group effects. Eval Rev 30(1):66–85

Pare P-P, Felson RB, Ouimet M (2007) Community variation in crime clearance multilevel analysis with comments on assessing police performance. J Quant Criminol 23(3):243–258

Rankin BH, Quane JM (2002) Social contexts and urban adolescent outcomes: the interrelated effects of neighborhoods, families, and peers on African-American youth. Soc Probl 49(1):79–100

Raudenbush SW (1984) Magnitude of teacher expectancy effects on pupil IQ as a function of credibility of expectancy induction: a synthesis of findings from 18 experiments. J Educ Psychol 76(1):85–97

Raudenbush SW (1989) "Centering" predictors in multilevel analysis. Choices and consequences. Multilevel Modeling Newsletter 1:10–12

Raudenbush SW, Bryk AS (2002) Hierarchical linear models: applications and data analysis methods, 2nd edn. SAGE Publications Inc

Robinson WS (1950) Ecological correlations and the behavior of individuals. Am Sociol Rev 15:351–357

Rosenfeld R, Fornango R, Rengifo AF (2007) The impact of order-maintenance policing in New York City homicide and robbery rates: 1988–2001. Criminology 45(2):355–384

Sampson RJ, Laub JH, Wimer C (2006) Does marriage reduce crime? A counterfactual approach to within-individual causal effects. Criminology 44(3):465–508

Schreck CJ, Stewart EA, Wayne Osgood D (2008) A reappraisal of the overlap of violent offenders and victims. Criminology 46(4):871–906

Simons RL, Simons LG, Burt CH, Brody GH, Cutrona C (2005) Collective efficacy, authoritative parenting and delinquency: a longitudinal test of a model integrating community- and felony-level processes. Criminology 43(4):989–1029

Slocum LA, Simpson S, Smith DA (2005) Strained lives and crime: examining intra-individual variation in strain and offending in a sample of incarcerated women. Criminology 43(4):1067–110

Smith D (1986) The neighborhood context of police behavior. Crime Justice 8:313–341

Snidjers T, Bosker R (1999) Multilevel models: an introduction to basic and advanced multilevel modeling. Sage Publications, Thousand Oaks

Stewart EA (2003) School social bonds, school climate, and school misbehavior: a multilevel analysis. Justice Q 20(3):575–604

Sun IY, Payne BK, Yuning W (2008) The impact of situational factors, officer characteristics, and neighborhood context on police behavior: a multilevel analysis. J Crim Justice 36(1):22–32

Taylor J, Malone S, Iacono WG, McGue M (2002) Development of substance dependence in two delinquency subgroups and nondelinquents from a male twin sample. J Am Acad Child Adolesc Psychiatry 41(4):386–393

Terrill W, Reisig MD (2003) Neighborhood context and police use of force. J Res Crime Delinq 40(3):291–321

Ulmer JT, Bradley MS (2006) Variation in trial penalties among serious violent offenses. Criminology 44(3):631–670

Ulmer JT, Johnson BD (2004) Sentencing in context: a multilevel analysis. Criminology 42(1):137–178

Warner BD (2007) Directly intervene or call the authorities? A study of forms of neighborhood social control within a social disorganization framework. Criminology 45(1):99–129

Wilcox P, Madensen TD, Tillyer MS (2007) Guardianship in context: implications for burglary victimization risk and prevention. Criminology 45(4):771–803

Wooldredge J, Griffin T, Pratt T (2001) Considering hierarchical models for research on inmate behavior: Predicting misconduct with multilevel data. Justice Q 18(1):203–231

Wooldredge J (2007) Neighborhood effects on felony sentencing. J Res Crime Delinq 44(2):238–263

Wright JP, Beaver KM (2005) Do parents matter in creating self-control in their children? A genetically informed test of Gottfredson and Hirschi's theory of low self-control. Criminology 43(4):1169–1202

Wright JP, Beaver KM, Delisi M, Vaughn M (2008) Evidence of negligible parenting influence on self-control, delinquent peers, and delinquency in a sample of twins. Justice Q 25(3):544–569

Wright DA, Bobashev G, Folsom R (2007) Understanding the relative influence of neighborhood, family, and youth on adolescent drug use. Subst Use Misuse 42:2159–2171

Wyant BR (2008) Multilevel impacts of perceived incivilities and perceptions of crime risk on fear of crime: isolating endogenous impacts. J Res Crime and Delinq 45(1):39–64

Xie M, McDowall D (2008) Escaping crime: the effects of direct and indirect victimization on moving. Criminology 46(4):809–840

Xie M, McDowall D (2008) The effects of residential turnover on household victimization. Criminology 46(3): 539–575

Zhang L, Messner SF, Liu J (2007) An exploration of the determinants of reporting crime to the police in the city of Tianjin, China. Criminology 45(4):959–984